**ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ**

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ


# Μεταπτυχιακή εργασία

## Μελέτη τεχνικών εξόρυξης δεδομένων και μηχανικής μάθησης για χρήση σε συστήματα ανίχνευσης εισβολών

Ευάγγελος Καραγιάννης

E-mail:vaggos1k@otenet.gr

A.M. MTE1513

Πειραιάς 2017

Επιβλέπων Καθηγητής: *Λαμπρινουδάκης Κωνσταντίνος*

*Η σελίδα αυτή είναι σκόπιμα λευκή.*

# Ευχαριστίες

*Θα ήθελα να ευχαριστήσω όλους όσους με βοήθησαν και με στήριξαν όλα αυτά τα χρόνια που πέρασα φοιτώντας στο Πανεπιστήμιο Πειραιώς. Επίσης θέλω να ευχαριστήσω τον επιβλέποντα καθηγητή Κωνσταντίνο Λαμπρινουδάκη και τον καθηγητή Βασίλειο Κάτο για τα εποικοδομητικά του σχόλια και τις συμβουλές του. Τέλος θέλω να ευχαριστήσω θερμά τον καθηγητή πληροφορικής και θείο μου Ηλία Γκρίνια που όλα αυτά τα χρόνια με βοηθούσε σε ότι και αν χρειάστηκα.*

# Πρόλογος

Τα τελευταία χρόνια παρουσιάζεται μια έξαρση στο αριθμό των επιθέσεων στο διαδίκτυο. Εκτός αυτού οι επιθέσεις πλέον είναι καλύτερα οργανωμένες και καταφέρνουν να διαπερνούν τα παραδοσιακά συστήματα προστασίας ενός πληροφοριακού συστήματος. Οι άνθρωποι που κρύβονται πίσω από αυτές τις επιθέσεις είναι πλέον επαγγελματίες με υψηλή κατάρτιση πάνω στο αντικείμενο. Βασικό κίνητρο τους είναι το κέρδος μιας και το ηλεκτρονικό έγκλημα έχει εξελιχθεί σε μια κερδοφόρα επιχείρηση πράγμα που σημαίνει ότι οι επιθέσεις θα εξελίσσονται συνεχώς και οι κυβερνο-εγκληματίες συνεχώς θα αναπτύσσουν νέες επιθέσεις για να διαπεράσουν τις ασπίδες προστασίας των πληροφοριακών συστημάτων. Αυτό έχει εγείρει αρκετές ανησυχίες στους ανθρώπους, τις εταιρίες και την επιστημονική κοινότητα που ασχολείται με την εύρεση νέων μεθόδων και ανάπτυξη νέων συστημάτων τα οποία θα είναι αποτελεσματικά και θα παρέχουν υψηλό επίπεδο ασφάλειας. Ιδιαίτερη προσοχή δίδεται στα συστήματα ανίχνευσης εισβολών *(Intrusion Detection Systems-IDS*) τα οποία είναι μια από τις βασικότερες ασπίδες άμυνας διότι είναι αυτά που θα πρέπει να ανιχνεύσουν τις απειλές την ώρα που πραγματοποιούνται, δηλαδή σε πραγματικό χρόνο. Τα συστήματα ανίχνευσης εισβολών είναι από τα κρισιμότερα κομμάτια του συνολικού μηχανισμού προστασίας διότι βρίσκεται στην πρώτη γραμμή άμυνας. Αν και τόσο κρίσιμο στοιχείο της όλης υποδομής τα συστήματα ανίχνευσης εισβολών χρησιμοποιούν αρκετά παλιές τεχνολογίες ανίχνευσης και βασίζονται κυρίως στην ανίχνευση εισβολών μέσω υπογραφών (*signature based*), μια μέθοδος που είναι ξεπερασμένη και πλέον αναποτελεσματική για την ανίχνευση εισβολών τύπου *zero-day* οι οποίες είναι η νούμερο ένα απειλή για ένα πληροφοριακό σύστημα. Για να αυξηθεί η αποτελεσματικότητα αυτών των συστημάτων οι επιστημονική έρευνα στράφηκε σε νέες μεθόδους και τεχνικές που χρησιμοποιούνται στους τομείς της εξόρυξης δεδομένων, της μηχανικής μάθησης και της ανίχνευσης ανωμαλιών. Αυτό συνέβη γιατί πλέον τα σύγχρονα συστήματα ανίχνευσης εισβολών θα πρέπει να είναι ικανά να διαχειρίζονται και να αναλύουν μεγάλους όγκους δεδομένων με σκοπό την ανίχνευση εισβολών και χωρίς συνεχή ανάδραση και παραμετροποίηση από τον χρήστη. Με λίγα λόγια θα πρέπει τα συστήματα να καταλαβαίνουν και να δρούνε από μόνα τους στις περισσότερες περιπτώσεις και ο χρήστης να εμπλέκεται μόνο για τις κρίσιμες αποφάσεις ή ενέργειες. Σε αυτήν την εργασία θα γίνει μελέτη και αναφορά κάποιων βασικών στοιχείων και μεθόδων όλων των παραπάνω τομέων που αναφέραμε, ιδιαίτερη προσοχή θα δοθεί στις τεχνικές μη επιβλεπόμενης (*unsupervised*) μηχανικής μάθησης και πιο συγκεκριμένα στην συσταδοποίηση υποχώρων (*subspace clustering*).

# Abstract

In recent years there has been an upsurge in the number of attacks on the internet. In addition, attacks are now better organised and manage to penetrate the traditional protection mechanisms of an information system. Highly trained professional are hiding behind these attacks. Their main motive is profit, since cybercrime has evolved into a profitable business, which means that attacks will evolve continuously and cyber criminals will constantly develop new attacks to penetrate the information protection shields. This has raised many concerns among people, companies and the scientific community that is concerned with finding new methods and developing new systems that will be effective and provide a high level of safety. Particular attention is paid to Intrusion Detection Systems (IDS), which are one of the most important defense shields because they are the ones that will have to detect the threats at the time they are, that is, in real time. Intrusion detection systems are among the most critical parts of the overall protection mechanism because they are at the forefront of defense. Although such a critical element of the whole infrastructure, intrusion detection systems use fairly old detection technologies and are mainly based on signature based signature detection, a method that is obsolete and no longer effective for detecting zero-day invasions that are the number one threat to an information system. To increase the effectiveness of these systems, scientific research has focused on new methods and techniques used in the fields of data mining, mechanical learning and anomalies detection. This is because modern intrusion detection systems now have to be able to handle and analyse large volumes of data in order to detect intrusions and without continuous feedback and customization by the user. In general, systems should be able to understand and act on their own in most cases, while the user should only be involved in critical decisions or actions. In this dissertation we will study and report some basic elements and methods of all the above-mentioned areas, special attention will be paid to unsupervised engineering techniques and more specifically to subspace clustering.

# Περιεχόμενα

# Κεφάλαιο 1ο

## Εισαγωγή

Την τελευταία εικοσαετία τα τεχνολογικά επιτεύγματα στον χώρο της πληροφορικής τόσο από την πλευρά του υλικού όσο και από την πλευρά του λογισμικού είναι γιγαντιαία σε σχέση με άλλους τομείς. Ο κλάδος της πληροφορικής είναι ένας από τους νεοσύστατους κλάδους και αυτήν την στιγμή βρίσκεται στον απόγειο της ανάπτυξης του. Χρόνο με τον χρόνο υλικό λογισμικό και δίκτυα αναπτύσσονται με γοργούς ρυθμούς. Νέα συστήματα παρουσιάζονται καθημερινά με σκοπό να επιλύσουν προβλήματα ή να βελτιώσουν παλιότερες ιδέες και λύσεις. Επιχειρήσεις και ιδιώτες χρησιμοποιούν τα αγαθά της πληροφορικής για να αυξήσουν και να βελτιώσουν την παραγωγικότητα της εργασίας του και όχι μόνο. Πλέον συσκευές όπως οι προσωπικοί υπολογιστές υπάρχουν σχεδόν σε κάθε σπίτι και οι έξυπνες συσκευές όπως τα έξυπνα τηλέφωνα (*smartphones*) υπάρχουν σε κάθε τσέπη. Χαρακτηριστικά σύμφωνα με το [**Error! Reference source not found.**], το έτος 2015 το περίπου το 44% του παγκόσμιου πληθυσμού έχει χρησιμοποιήσει το διαδίκτυο, ενώ πέντε χρόνια πριν δηλαδή το 2010 το ποσοστό αυτό ανερχόταν στο 29%. Από τα στατιστικά λοιπόν καταλαβαίνουμε ότι η αύξηση του αριθμού των ανθρώπων που χρησιμοποιούν των διαδίκτυο σε παγκόσμια κλίμακα είναι ραγδαία. Οι δυνατότητες και οι ευκολίες που μας προσφέρει το διαδικτύου και τα πληροφοριακά συστήματα είναι αναμφισβήτητες και για αυτό η χρήση του αυξάνεται όλο και περισσότερο, τόσο για επαγγελματικούς σκοπούς όσο και για ψυχαγωγικούς.

Η πληροφορική και το διαδίκτυο είναι αξιοσημείωτα επιτεύγματα του ανθρώπινου πολιτισμού και έδωσε λύσεις σε πολλά προβλήματα της καθημερινότητας και της ανθρωπότητας γενικότερα. Ωστόσο ένα ποσοστό χρηστών αυτών των συστημάτων έχει διαφορετικά κίνητρα από την απλή χρήση ή ανάπτυξη νέων λύσεων. Έτσι από τα πρώτα χρόνια της πληροφορικής εμφανίστηκαν κάποιοι άνθρωποι, οι οποίοι συνήθως ήταν άτομα με πολύ καλή τεχνική γνώση του τρόπου λειτουργίας αυτών των συστημάτων και τα οποία προσπάθησαν να «πειράξουν» υλικό ή λογισμικό επηρεάζοντας την ορθή λειτουργία των συστημάτων. Αυτοί οι άνθρωποι είναι οι κοινώς αποκαλούμενοι «χάκερς». Τα πρώτα χρόνια το κυρίως κίνητρο των χάκερ ήταν η εξερεύνηση του τρόπου λειτουργίας των συστημάτων και οι οποιοσδήποτε μετατροπές που πραγματοποιήσαν είχαν ως κύριο κίνητρο το να προκαλέσουν προβλήματα στους χρήστες των πληροφοριακών συστημάτων. Όμως πολύ σύντομα εμφανίστηκαν και άλλα κρούσματα επιθέσεων με σκοπό την υποκλοπή δεδομένων, παραβίαση ή παρεμπόδιση ορθής λειτουργίας των πληροφοριακών συστημάτων. Στις μέρες μας οι «χάκερς» πλέον δεν δρούνε αυτόνομα αλλά κυρίως ως ομάδες. Σκοπός τους πλέον δεν είναι απλά να προκαλέσουν μόνο προβλήματα και να παραβιάσουν συστήματα αλλά το κέρδος [**Error! Reference source not found.**][**Error! Reference source not found.**]. Μέσω διαφορετικών μορφών επιθέσεων (ιοί , *ransomware* κ.λ.π ) που εξαπολύουν έχουν την δυνατότητα να διαπεράσουν τις όποιες ασπίδες προστασίας υπάρχουν στα σύγχρονα πληροφορικά συστήματα να αποσπάσουν πολύτιμες πληροφορίες από αυτά η ακόμα και να πάρουν τον έλεγχο των συστημάτων προκαλώντας ανυπολόγιστες ζημιές. Αυτοί η σύγχρονη μορφή εγκλήματος αποκαλούμενο και ως κύβερνο-έγκλημα, προβληματίζει τους κατόχους πληροφορικών υποδομών αλλά και τις εταιρείες ανάπτυξης προϊόντων (*antivirus, IDS, firewall* κ.λ.π) προστασίας τέτοιων συστημάτων.

Αν και υπάρχει μεγάλη εξέλιξη στα προϊόντα και τις λύσεις που προσφέρονται για την προστασία μιας πληροφοριακής υποδομής, οι επιτιθέμενοι βρίσκουνε διαρκώς νέους τρόπους και αναπτύσσουν νέες έξυπνες επιθέσεις έτσι ώστε να διαπεράσουν τις υπάρχουσες ασπίδες ασφάλειας. Εκτός αυτού ο

αριθμός των επιθέσεων αυξάνεται ολοένα και περισσότερο κάθε χρόνο[6]. Επιθέσεις τύπου *Distributed Denial of Service* (*DDoS*) , ιοί και κακόβουλα λογισμικά (*malwares*) είναι κάποιες από τις επιθέσεις με τις οποίες έρχονται αντιμέτωπα τα πληροφορικά συστήματα καθημερινά. Λόγω αυτής της ραγδαίας αύξησης των επιθέσεων και των εξελιγμένων μορφών επιθέσεων η επιστημονική κοινότητα που ασχολείται με την προστασία πληροφοριακών συστημάτων τα τελευταία χρόνια καταβάλει μεγάλη προσπάθεια στην ανάπτυξη νέων τεχνικών και μεθόδων που θα μπορούνε να αντιμετωπίσουν αποτελεσματικά το μεγαλύτερο ποσοστό αυτών των επιθέσεων. Νέου τύπου συστήματα μελετώνται τα οποία προσεγγίζουν το πρόβλημα διαφορετικά από ότι το αντιμετώπιζαν κάποια χρόνια πριν. Ένα από αυτά τα συστήματα είναι και τα συστήματα ανίχνευσης εισβολών (*Intrusion Detection Systems*). Τα συστήματα ανίχνευσης εισβολών έχουν ως κύριο σκοπό την ανίχνευση και αποτροπή μιας επίθεσης σε πραγματικό χρόνο, δηλαδή την ώρα που συμβαίνει. Τα υπάρχοντα συστήματα που κυκλοφορούν στο εμπόριο και είναι εγκατεστημένα στους περισσότερους οργανισμούς αποτυγχάνουν σε μεγάλο βαθμό να αντιμετωπίσουν εκλεπτυσμένες και πρωτοεμφανιζόμενες επιθέσεις τύπου *zero day* . Επιπλέον χρειάζονται συνεχή παραμετροποίηση και ανάδραση από τον διαχειριστή του συστήματος έτσι ώστε να είναι ενήμερα για την κάθε είδους νέα επίθεση που εμφανίζεται. Ένα ακόμα πρόβλημα που παρουσιάζεται είναι ότι συνήθως δεν μπορούν να αναλύσουν μεγάλο όγκο δεδομένων ο οποίος προέρχεται κατά κύριο λόγο από την διαδικτυακή κίνηση.

## 1.1 Σκοπός της εργασίας

Τα όσα ειπώθηκαν παραπάνω αποτέλεσαν και το κίνητρο αυτής της διπλωματικής. Στόχος αυτής της διπλωματικής εργασίας είναι η βιβλιογραφική διερεύνηση νέων τεχνικών οι οποίες θα μπορέσουν να εφαρμοστούν στα συστήματα ανίχνευσης εισβολών έτσι ώστε να μπορούν να ανιχνεύσουν γνωστές αλλά και πρωτοεμφανιζόμενες επιθέσεις με όσο το δυνατό λιγότερη ανάδραση από τον διαχειριστή αυτού του συστήματος. Για την εύρεση πιθανών λύσεων στα παραπάνω προβλήματα η έρευνα μας στράφηκε στους τομείς της εξόρυξης δεδομένων (*Data Mining*) έτσι ώστε να μελετήσουμε τεχνικές οι οποίες μπορούν να διαχειριστούν μεγάλους όγκους δεδομένων, στην μηχανική μάθηση η οποία προσφέρει λύσεις αυτοματοποίησης συστημάτων μειώνοντας την ανάδραση των χρηστών στο ελάχιστο δυνατό και τέλος στον τομέα της ανίχνευσης ανωμαλιών ο οποίος εκμεταλλευόμενος τεχνικές των δύο προηγούμενων τομέων προτείνει διάφορες προσεγγίσεις με σκοπό την ανίχνευση μη ομαλών-ύποπτων συμπεριφορών, που στην προκείμενη περίπτωση αυτές οι μη ομαλές συμπεριφορές είναι οι επιθέσεις που θέλουμε να ανιχνεύσουμε.

## 1.2 Δομή της εργασίας

Στο κεφάλαιο 2 αναφέρουμε τι είναι ένα σύστημα ανίχνευσης εισβολών, κάποιες βασικές κατηγορίες τέτοιων συστημάτων και τις τεχνολογίες ανίχνευσης που χρησιμοποιούν. Αναφορά γίνεται και στις κατηγορίες επιθέσεων που θα πρέπει να αντιμετωπίζουν αυτά τα συστήματα για να ξεκαθαριστεί ο σκοπός αυτών των συστημάτων ο οποίος είναι κυρίως η ανίχνευση και αποτροπή συγκεκριμένων επιθέσεων. Στο τέλος αυτού του κεφαλαίου αναφέρουμε και κάποια μέτρα με τα οποία αξιολογούμε την αποτελεσματικότητα ενός συστήματος ανίχνευσης εισβολών.

Στο κεφάλαιο 3 γίνεται μια εκτενέστερη αναφορά στην Εξόρυξη Δεδομένων (*Data Mining*) και στις τεχνικές και μεθόδους που χρησιμοποιούνται για εξόρυξη δεδομένων. Επιπρόσθετα αναφερόμαστε στα προβλήματα που αντιμετωπίζουμε κατά την προσπάθεια εξαγωγής πληροφορίας από μεγάλους όγκους δεδομένων και τέλος αναφερόμαστε σε κάποιες έρευνες που έχουν δημοσιευτεί και σχετίζονται με την προσπάθεια εφαρμογής τεχνικών εξόρυξης δεδομένων για την υλοποίηση συστημάτων ανίχνευσης εισβολών.

Στο κεφάλαιο 4 περιγράφεται η Μηχανική Μάθηση η οποία είναι ένας δημοφιλής κλάδος της πληροφορικής και γίνεται όλο και πιο δημοφιλής διότι οι τεχνικές που προσφέρει έχουν βελτιστοποιηθεί σε τέτοιο βαθμό όπου πλέον μπορούμε να τις χρησιμοποιήσουμε για την επίλυση προβλημάτων που απασχολούσαν δεκαετίες την ανθρωπότητα. Ειδικά στον τομέα της ασφάλειας των υπολογιστών η μηχανική μάθηση προβλέπεται να παίζει κυρίαρχο ρόλο τα επόμενα χρόνια για την ανάπτυξη συστημάτων τα οποία θα λαμβάνουν από μόνα τους επιφάσεις και θα δρουν ανάλογα σε κάθε περίσταση σύμφωνα πάντα με τις εντολές που τους έχουν δοθεί από τους ανθρώπους. Σε αυτό το κεφάλαιο περιγράφονται τα είδη της μηχανικής μάθησης και οι αντίστοιχοι αλγόριθμοι που υπάρχουν, δίνοντας μεγαλύτερη έμφαση στην μη επιβλεπόμενη (*unsupervised*) μάθηση.

Στο κεφάλαιο 5 γίνετε μια εκτενής αναφορά στην τεχνική της ανίχνευσης ανωμαλιών. Στην αρχή του κεφαλαίου γίνεται αναφορά στα είδη των επιθέσεων έτσι ώστε να γίνει κατανοητό το δύναται να αντιμετωπίζει ένα σύστημα ανίχνευσης εισβολών. Το υπόλοιπο του κεφαλαίου αναφέρεται σε τεχνικές και μεθόδους οι οποίες έχουν μελετηθεί κατά καιρούς και οι οποίες εφαρμόζονται για την ανίχνευση ακραίων τιμών (*outliers*), τα οποία υποδηλώνουν στην πλειοψηφία των περιπτώσεων μη ομαλή συμπεριφορά η οποία συνήθως υποδηλώνει εισβολή.

Στο κεφάλαιο 6 γίνεται αναφορά στα πολυδιάστατα δεδομένα, την επιλογή χαρακτηριστικών, την συσταδοποίηση (*clustering*) και την συσταδοποίηση υποχώρων (*subspace clustering*). Η συσταδοποίηση υποχώρων είναι η τεχνική με την οποία μπορούμε να διαχειριστούμε ένα μεγάλο όγκο δεδομένων ο οποίος απαρτίζεται και από πολλές διαστάσεις. Αυτό καθιστά αυτήν την τεχνική κατάλληλη για εφαρμογή σε συστήματα ανίχνευσης εισβολών νέας γενιάς. Στο συγκεκριμένο κεφάλαιο γίνεται λεπτομερής αναφορά σε όλες τις μεθόδους και τους αλγορίθμους που βασίζουν τον τρόπο λειτουργίας τους στην τεχνική της συσταδοποίησης υποχώρων.

# Κεφάλαιο 2ο

## Συστήματα Ανίχνευσης Εισβολών (*Intrusion Detection Systems*)

Όπως είναι γνωστό οι υπολογιστές και γενικότερα τα πληροφοριακά συστήματα γίνονται στόχος επιτιθέμενων οι οποίοι έχουν ποικίλους σκοπούς και πλέον τα τελευταία χρόνια το οργανωμένο έγκλημα στο διαδίκτυο βρίσκεται στην άνθισή του. Για αυτό το λόγο έχουν αναπτυχθεί διαφορά συστήματα για την προστασία των πληροφοριακών υποδομών και ένα από αυτά είναι και το συστήματα ανίχνευσης εισβολών (*Intrusion Detection Systems* σε συντομογραφία *IDS*). Το σύστημα ανίχνευσης εισβολών είναι πλέον αναπόσπαστο κομμάτι κάθε ασφαλούς πληροφοριακού συστήματος και παίζει σημαντικό ρόλο στην διαφύλαξη του και την ανίχνευση επιθέσεων διαφόρων ειδών που αυτό δέχεται.

Το πρώτο μοντέλο για *IDS* προτάθηκε από τον *Denning* το 1987 [4] και αναφερόταν σε ένα σύστημα *IDS* πραγματικού χρόνου το οποίο θα βοηθούσε στην αυτόματη ανίχνευση εισβολών και προσπαθειών εισβολών σε ένα πληροφοριακό σύστημα. Επίσης το προτεινόμενο μοντέλο ήταν γενικού σκοπού και μπορούσε να εφαρμοστεί σε όλα τα πληροφοριακά συστήματα διότι είχε αναπτυχθεί για αυτόν τον σκοπό. Δηλαδή κατασκευάστηκε έτσι ώστε να έχει καθολική εφαρμογή παντού και να είναι ανεξάρτητο από συγκεκριμένα συστήματα, περιβάλλον εφαρμογής, είδος επίθεσης και ευπάθειας συστήματος. Όπως καταλαβαίνουμε αυτό το μοντέλο ήταν το έναυσμα για μια σειρά ερευνών οι οποίες έχουν ως κύριο στόχο να κατασκευάσουν μοντέλα που είναι αποδοτικά και ακριβή. Στα τέλη της δεκαετίας του 1980 και μέχρι τις αρχές του 1990 ο συνδυασμός στατιστικών προσεγγίσεων (*statistical approaches*) και ειδικών συστημάτων (*expert systems*) μονοπωλούσαν το ενδιαφέρον των ερευνητών [5]. Από τα μέσα του 1990 μέχρι και σήμερα η προσοχή των ερευνητών στρέφεται πλέον σε τεχνικές τεχνητής νοημοσύνης (*artificial intelligence*) και μηχανικής μάθησης (*machine learning*).

Σε αυτό το κεφάλαιο θα περιγράψουμε περιληπτικά κάποια βασικά δομικά στοιχεία των *IDS* όπως τις κατηγορίες αυτών , τις αρχιτεκτονικές τους, την μεθοδολογία που χρησιμοποιούν για ανίχνευση εισβολών, την προσέγγιση που χρησιμοποιούν για την ανίχνευση εισβολών, τα είδη των επιθέσεων που αντιμετωπίζουν και τέλος κάποιες μετρήσεις με τις οποίες εξετάζουμε την αποτελεσματικότητα και την αποδοτικότητα του *IDS*.

### 2.1 Κατηγορίες IDS

Τα συστήματα ανίχνευσης εισβολών (*IDS)* χωρίζονται σε δύο βασικές κατηγορίες όσων αφορά τον τρόπο εφαρμογή τους: Τα *Host Based IDS (HIDS)* και τα *Network Based IDS (NIDS)*. Υπάρχει και μια τρίτη κατηγορία η οποία είναι συνδυασμός των δύο παραπάνω και ονομάζονται *Hybrid (HIDS)*.

- Host Based IDS-HIDS: Σε αυτήν την κατηγορία το *IDS* σύστημα εγκαθίσταται σε ένα υπολογιστή και η ανίχνευση των εισβολών λαμβάνει χώρα μόνο για αυτό το μηχάνημα. Αν θέλουμε προστασία πολλών συστημάτων τότε το *HIDS* περιλαμβάνει προγράμματα τα αποκαλούνται πράκτορες (*agents*) και τα οποία εγκαθίστανται σε κάθε φυσικό μηχάνημα ξεχωριστά, έτσι ώστε να συλλέγουν πληροφορίες(*log files ,system calls, network activity*) για αυτό το μηχάνημα και την λειτουργία του. Συνήθως στις περισσότερες υλοποιήσεις αυτού του τύπου υπάρχει ένας κεντρικός υπολογιστής *server* ο οποίος συλλέγει τα δεδομένα αυτά για ευκολότερη ανάλυση και εξέταση. Έπειτα αν υπάρχει κάποια μη εξουσιοδοτημένη αλλαγή ή δραστηριότητα σε ένα από

αυτά τα αρχεία που παρακολουθεί και ελέγχει τότε δημιουργεί κάποιο συναγερμό έτσι ώστε να ενημερώσει το χρήστη ότι κάτι συμβαίνει στο σύστημα και να μπλοκάρει την συγκεκριμένη ενέργεια που ανίχνευσε ως μη εξουσιοδοτημένη, για να μην υπάρξει περαιτέρω ζημία στο σύστημα[6][7].

- Network Based IDS-NIDS: Αυτού του είδους *IDS* χρησιμοποιούνται για να παρακολουθήσουν το δίκτυο και να αναλύσουν την δικτυακή κίνηση σε επίπεδο *bytes*, πακέτου ή δικτυακής ροής (*network flow*) έτσι ώστε να προστατεύσουν ένα σύστημα από επιθέσεις που προέρχονται από το δίκτυο [8]. Τέτοιου είδους επιθέσεις χαρακτηριστικά είναι οι *Denial of Service (DoS)* και οι *Distributed Denial of Service (DDoS)* επιθέσεις οι οποίες αποτελούν την μεγαλύτερη απειλή στις για τα πληροφοριακό σύστημα κυρίως μεγάλων οργανισμών στις μέρες μας [9]. Τα *NIDS* σαν δομή απαρτίζονται από πολλούς αισθητήρες οι οποίοι τοποθετούνται σε διαφορά σημεία του δικτύου του πληροφοριακού συστήματος για να παρακολουθούν την δικτυακή κίνηση και έναν ή περισσότερους *server* για τον έλεγχο των αισθητήρων αυτών και την συλλογή δεδομένων από αυτούς [6]. Όσον αφορά τα *NIDS* επειδή έχουν να επεξεργαστούν μεγάλο όγκο δεδομένων και πολλά είδη επιθέσεων οι έρευνες τα τελευταία χρόνια στρέφονται σε *NIDS* τα οποία εφαρμόζουν τεχνικές ανίχνευσης ανωμαλιών (*anomaly based*) [10][11] και δεν χρειάζονται επίβλεψη (*unsupervised*) έτσι ώστε να εντοπίσουν γνωστές και άγνωστες επιθέσεις (*zero-day*) [12][13][14].

- Hybrid IDS-HIDS: Τα υβριδικά (*hybrid*) *IDS* είναι ο συνδυασμός των δύο παραπάνω κατηγοριών *IDS*. Είναι η μια κατηγορία *IDS* που παρέχει μεγάλη ασφάλεια καλύπτοντας το σύνολο ενός πληροφοριακού συστήματος διότι ελέγχει δίκτυο και μηχανήματα [6][15].

## 2.2 Αρχιτεκτονικές IDS

Σε αυτήν την ενότητα θα αναφέρουμε τις βασικές αρχιτεκτονικές στις οποίες τα *IDS* χωρίζονται βάσει της αρχιτεκτονικής με την οποία επικοινωνούν. Οι αρχιτεκτονικές αυτές είναι η κεντρική (*centralized*), η αποκεντρωμένη (*decentralized*) και η διανεμημένη (*distributed*).

- Κεντρική (*Centralized*) : Σε αυτήν την αρχιτεκτονική το *IDS* αποτελείται από αρκετά προγράμματα πράκτορες (*agents*) στα μηχανήματα και αισθητήρες (*sensor*) σε διάφορα σημεία του δικτύου. Όλες οι πληροφορίες που συλλέγονται από τα παραπάνω μεταφέρονται σε μια κεντρική μονάδα επεξεργασίας και ανάλυσης η οποία τα επεξεργάζεται και δημιουργεί αντίστοιχα συναγερμούς(*alerts*) αν έχει ανιχνευθεί κάποια επίθεση. Αυτός ο τρόπος λειτουργίας και επεξεργασίας είναι παθητικός και δεν είναι ικανός να προστατέψει μεγάλες πληροφοριακές υποδομές και υποδομές μου συνεχώς προστίθενται νέα κομμάτια στο πληροφοριακό τους σύστημα [16]. Επιπλέον σε περίπτωση μεγάλου φόρτου εργασίας η κεντρική μονάδα επεξεργασίας και ανάλυσης μπορεί να μην μπορεί να ανταπεξέλθει στον μεγάλο φόρτο εργασίας και αυτό είναι κρίσιμο για ένα σύστημα *IDS* διότι πρέπει να λειτουργεί και να ανταποκρίνεται σε πραγματικό χρόνο. Επίσης στην κεντρική αρχιτεκτονική η κεντρική μονάδα επεξεργασίας και ανάλυσης είναι ένα *Single Point of Failure (SPoF)*, δηλαδή αν σταματήσει να δουλεύει ή αργήσει να επεξεργαστεί τα δεδομένα το σύστημα μένει εκτιθέμενο [17].

- Αποκεντρωμένη (*Decentralized*): Εδώ το *IDS* αποτελείται πάλι από πολλούς πράκτορες (*agents*) οι οποίοι είναι εγκατεστημένοι σε διαφορά κομβικά σημεία του πληροφοριακού συστήματος οι

οποίοι αναλαμβάνουν να προ-επεξεργαστούν τα δεδομένα που παρακολουθούν, να ανιχνεύσουν τοπικά επιθέσεις και να συνεργαστούν με τους υπόλοιπους πράκτορες στο δίκτυο. Σε αυτήν την αρχιτεκτονική λόγω της προ-επεξεργασίας που γίνεται στους πράκτορες αποφεύγετε η υπολογιστική επιβάρυνση της κεντρικής μονάδας επεξεργασίας και ελέγχου.

- Διανεμημένη (*Distributed*): Εδώ δεν υπάρχει κεντρική μονάδα επεξεργασίας και ανάλυσης Ένα *Distributed IDS* μπορεί να χαρακτηριστεί ως ένα σύνολο μεμονωμένων *IDS* σε ένα μεγάλο δίκτυο τα οποία επικοινωνούν μεταξύ τους [16] και διαμοιράζονται τον φόρτο εργασίας που θα έκανε η κεντρική μονάδα επεξεργασίας και ανάλυσης επειδή κάθε πράκτορας(*agent*) λειτουργεί σαν αυτόνομο *IDS*. Συνήθως τα *Distributed IDS* επικοινωνούν με μεταξύ τους χρησιμοποιώντας έναν *Peer-to-Peer(P2P)* τρόπο επικοινωνίας [17]. Τέτοιου είδους *IDS* συναντούμε κυρίως σε κινητά *ad hoc* δίκτυα [18] και σε δίκτυα ασύρματων αισθητήρων [19].

## 2.3 Μεθοδολογίες Ανίχνευσης

Οι μεθοδολογίες ανίχνευσης των συστημάτων ανίχνευσης εισβολών χωρίζονται σε τρεις κατηγορίες. Η πρώτη κατηγορία βασίζεται στην ανίχνευση μέσω υπογραφών (*Signature-based Detection*) , η δεύτερη στην ανίχνευση ανωμαλιών (*Anomaly-based Detection*) και η τρίτη στην ανάλυση *stateful* πρωτοκόλλων (*Stateful Protocol Analysis*). Παρακάτω θα περιγράψουμε περιληπτικά κάποια βασικά στοιχεία αυτών των μεθόδων.

- Ανίχνευση μέσω υπογραφών (*Signature-based Detection*): Μια υπογραφή (*signature*) είναι ένα συγκεκριμένο υπόδειγμα, μια μοναδική ταυτότητα που έχει δοθεί σε μια ήδη γνωστή επίθεση. Το *IDS* έχει στην βάσει δεδομένων του υπογραφές όλων των γνωστών επιθέσεων που έχουν καταγραφεί και χρησιμοποιεί αυτές τις υπογραφές για τις συγκρίνει με περιστατικά (*events*) που έχουν συμβεί και να αποφανθεί αν αυτά τα περιστατικά είναι επιθέσεις εισβολής. Το δύο μεγάλα μειονεκτήματα αυτής της μεθόδου είναι, πρώτον ότι μόνο γνωστές απειλές ανιχνεύονται και δεύτερον κάθε καινούργια απειλή που εντοπίζεται πρέπει να αναλυθεί από τους ειδικούς και να δημιουργεί μια υπογραφή για αυτήν την απειλή έτσι ώστε να γίνει αναβάθμιση της βάσεις δεδομένων και να μπορεί στο μέλλον το *IDS* να αναγνωρίσει αυτήν την απειλή [20].

- *Ανίχνευση ανωμαλιών (Anomaly-based Detection)* : Η διαδικασία της ανίχνευση ανωμαλιών βασίζεται στην λογική της σύγκρισης μιας οποιασδήποτε δραστηριότητας με μια φυσιολογική δραστηριότητα και την εύρεση των διαφορών μεταξύ αυτών των δύο. Αν η εξεταζόμενη δραστηριότητα έχει αρκετές διαφορές με την φυσιολογική δραστηριότητα τότε η πρώτη είναι χαρακτηρίζεται ως εισβολή [21]. Για να μπορέσει αυτήν η τεχνική να λειτουργήσει αρχικά το *IDS* θα πρέπει να καταγράψει δραστηριότητες του συστήματος, του δικτύου και των χρηστών για ένα χρονικό διάστημα. Με αυτόν τον τρόπο θα δημιουργήσει κάποια προφίλ φυσιολογικής συμπεριφοράς για τον τρόπο λειτουργίας του συστήματος. Έπειτα θα μπορεί να συγκρίνει αυτά τα προφίλ με γεγονότα που παρατηρεί για να αποφανθεί αν αυτά είναι απειλές ή όχι [22]. Το μεγαλύτερο πρόβλημα αυτής της μεθόδου είναι ότι υπάρχει μεγάλος αριθμός *false positives*, δηλαδή φυσιολογικές δραστηριότητες ανιχνεύονται ως εισβολές. Αυτό συμβαίνει συνήθως σε δυναμικά περιβάλλοντα που οι δραστηριότητες του συστήματος μεταβάλλονται συνεχώς [23].

- Ανάλυση *stateful* πρωτοκόλλων (*Stateful Protocol Analysis*): Η λέξη *stateful* σε αυτήν την μεθοδολογία σημαίνει ότι το *IDS* είναι ικανό να καταλάβει και να εντοπίσει την κατάσταση (*state*)

των πρωτοκόλλων δικτύου, μεταφοράς και εφαρμογής, των οποίων γνωρίζει εκ των υστέρων τον τρόπο λειτουργίας τους . Σε αντίθεση με την *Anomaly-based* μέθοδο εδώ αντί το *IDS* να φτιάχνει προφίλ για την λειτουργία του πληροφοριακού συστήματος, βασίζεται σε προφίλ πρωτοκόλλων που αναπτύσσονται από συγκεκριμένους προμηθευτές λογισμικού, για συγκεκριμένα πρωτόκολλα, αλλά και σε αυτά των διαφόρων οργανισμών όπως ο *Internet Engineering Task Force (IETF)* και *Request for Comments (RFC)* , τα οποία προφίλ περιγράφουν πως τα πρωτόκολλα λειτουργούν [22].

Αναλυτικότερα τα πλεονεκτήματα και τα μειονεκτήματα των παραπάνω μεθοδολογιών περιγράφονται στα [22][23].

## 2.4 Μέθοδοι Ανίχνευσης Εισβολών

Ένα σύστημα ανίχνευσης εισβολών (*IDS)* παρακολουθεί την δραστηριότητα ενός πληροφοριακού συστήματος και προσπαθεί να εντοπίσει από τις διεργασίες ή δραστηριότητα του δικτύου, ασυνήθιστες συμπεριφορές οι οποίες να υποδεικνύουν προσπάθειες επίθεσης προς το πληροφοριακό σύστημα. Όσων αφορά τις μεθόδους ανίχνευσης εισβολών τα *IDS* έχουν δύο κατηγορίες την ανίχνευση κακής χρήσης (*misuse detection*) και την ανίχνευση ανωμαλιών (*anomaly detection*).

- Ανίχνευση κακής χρήσης (*misuse detection*): Με αυτήν την τεχνική το *IDS* αναλύει όλη την πληροφορία που συλλέγει και προσπαθεί να την συγκρίνει με υπογραφές (“*signatures*”) επιθέσεων που έχει στην βάσει δεδομένων του [24]. Με αυτήν την μέθοδο τον *IDS* μπορεί να ανιχνεύσει μόνο γνωστές απειλές οι οποίες έχουν ήδη αναλυθεί και έχουν μια ψηφιακή υπογραφή με μεγάλη επιτυχία. Αυτήν η τεχνική εφαρμόζεται στα περισσότερα εμπορικά συστήματα διότι είναι αξιόπιστη και έχει καλά αποτελέσματα και χαμηλό αριθμό *false positives* [5]. Από την άλλη πλευρά σε καινούργιές μορφές απειλών που δεν έχουν ακόμα αναλυθεί και δεν έχουν υπογραφή αυτήν η τεχνική αποτυγχάνει πλήρως.

- Ανίχνευση ανωμαλιών (*anomaly detection*): Αυτήν η τεχνική έχει διαφορετική προσέγγιση από την προηγούμενη και βασίζεται στην παρακολούθηση της συμπεριφοράς του συστήματος. Αν η συμπεριφορά του συστήματος κάποια στιγμή είναι διαφορετική από ότι συνήθως τότε το σύστημα εγείρει συναγερμό για πιθανή εισβολή. Για να καθορίσουμε την φυσιολογική συμπεριφορά ενός συστήματος το *IDS* πρέπει να έχει κάποια προφίλ φυσιολογικής συμπεριφοράς. Αυτά τα προφίλ πρέπει να δημιουργηθούν από του ειδικούς που στήνουν το σύστημα και να ανανεώνονται αρκετά συχνά έτσι ώστε το *IDS* να είναι ενημερωμένο συνεχώς για τις διάφορες αλλαγές που συμβαίνουν [25]. Ειδικότερα σε δυναμικά περιβάλλοντα που οι αλλαγές στην κίνηση του δικτύου για παράδειγμα μεταβάλλεται συνεχώς τα προφίλ φυσιολογικής συμπεριφοράς πρέπει να αλλάζουν σε τακτά χρονικά διαστήματα. Η συνεχής αλλαγή του προφίλ φυσιολογικής συμπεριφοράς αλλά και η δυσκολία να καθοριστεί τι είναι φυσιολογική συμπεριφορά και τι όχι είναι τα μεγαλύτερα μειονεκτήματα αυτής της τεχνικής [15]. Τέλος υπάρχουν δύο είδη ανίχνευσης ανωμαλιών. Η πρώτη ονομάζεται στατική ανίχνευση ανωμαλιών (*static anomaly detection*) η οποία υποθέτει η συμπεριφορά του συστήματος δεν αλλάζει ποτέ (π.χ. *system calls*) και η δεύτερη ονομάζεται δυναμική ανίχνευση ανωμαλιών (*dynamic anomaly detection*) η οποία δημιουργεί προφίλ για τις συνήθειες των χρηστών, της χρήσης του δικτύου και των μηχανήματων [26].

## 2.5 Κατηγορίες επιθέσεων που αντιμετωπίζουν τα IDS

Οι κατηγορίες επιθέσεων γενικά που πρέπει να αντιμετωπίσει ένα *IDS* είναι τέσσερις και είναι οι εξής:

- Denial of Service (*DoS*)/Distributed Denial of Service (DDoS): Στην *DoS* επίθεση ο επιτιθέμενος δημιουργεί υπολογιστικό φόρτο σε ένα σύστημα στέλνοντας συνεχώς αιτήσεις (π.χ *ping flood*) από μια πηγή, με σκοπό να εξαντλήσει τους πόρους του συστήματος και να θέσει το σύστημα μη διαθέσιμο στους υπόλοιπους χρήστες που το χρειάζονται να το χρησιμοποιήσουν. Η *DDoS* έχει την ίδια φιλοσοφία με την *DoS* απλά ο επιτιθέμενος χρησιμοποιεί πλέον πολλές πηγές από τις οποίες εξαπολύει την επίθεση του [27].
- User to Root Attack (*U2R*): Αυτήν η επίθεση αναφέρεται σε επιτιθέμενους που είναι κανονικοί χρήστες και θέλουν να αποκτήσουν δικαιώματα πρόσβασης υπέρ-χρήστη (*root*) σε ένα σύστημα εκμεταλλευόμενοι μια ευπάθεια του συστήματος [28].
- Remote to Local Attack (*R2L*): Είναι η επίθεση στην οποία κάποιος χρήστης καταφέρνει να αποκτήσει μη εξουσιοδοτημένη πρόσβαση σε ένα σύστημα απομακρυσμένα, δηλαδή χωρίς να χρειάζεται να βρίσκεται γεωγραφικά στον ίδιο χώρο όπου βρίσκεται το σύστημα στόχος [25].
- Probing attack (*PROB*): Σε αυτού του είδους την επίθεση ο επιτιθέμενος προσπαθεί να συλλέξει πληροφορίες για το σύστημα στόχο χωρίς να έχει δικαιώματα πρόσβασης με σκοπό να βρει ευπάθειες που έπειτα θα τις εκμεταλλευτεί για να αποκτήσει πρόσβαση σε αυτό. Σε αυτές τις περιπτώσεις ο επιτιθέμενος χρησιμοποιεί τεχνικές *port scanning, ping sweep* κ.λ.π [28][29].

## 2.6 Μέτρα αποτελεσματικότητας και αποδοτικότητας των IDS

Ένα σύστημα ανίχνευσης εισβολών IDS πρέπει να τηρεί κάποιες προϋποθέσεις οι οποίες πρέπει να ληφθούν υπόψη κατά την δημιουργία του. Όπως αναφέρεται και στο [**Error! Reference source not found.**] ένα *IDS* πρέπει να αξιολογηθεί βάσει πέντε μετρήσεων αποδοτικότητας. Τα πέντε αυτά μετρά είναι τα εξής:

- Ακρίβεια (*Accurancy*): Ένα *IDS* πρέπει να είναι ακριβές στις προβλέψεις του και να αναγνωρίζει με όσο το δυνατόν μεγαλύτερη ακρίβεια τις κακόβουλες επιθέσεις. Από την άλλη ένα *IDS* θεωρείται ανακριβές όταν ανιχνεύει εξουσιοδοτημένες ενέργειες ως εισβολές ή δεν αναγνωρίζει κάποιες εισβολές.
- Απόδοση (*Performance*): Η απόδοση ενός *IDS* αξιολογείται από τον ρυθμό με τον οποίο επεξεργάζεται τα διάφορα γεγονότα που υπόκεινται στον έλεγχο του (*audit events*). Αν το *IDS* επεξεργάζεται αυτά τα γεγονότα με αργό ρυθμό τότε η ανίχνευση σε πραγματικό χρόνο είναι αδύνατη.
- Πληρότητα (*Completeness*): Ένα *IDS* δεν είναι πλήρες όταν αποτυγχάνει να ανιχνεύσει κάποια επίθεση. Όπως καταλαβαίνουμε αυτό το μέτρο αποδοτικότητας είναι δύσκολα μετρήσιμο διότι αδύνατο να έχουμε πλήρη γνώση για τον αριθμό των διαδικτυακών επιθέσεων που υπάρχουν σε παγκόσμιο επίπεδο. Αν λάβουμε υπόψη και τις καινούργιες επιθέσεις που εμφανίζονται συνεχώς, τις ευπάθειες που έχουν βρεθεί και δεν έχουν δημοσιευτεί, τις *zero-days* επιθέσεις κ.λ.π ,τότε καταλαβαίνουμε ότι η έννοια της πληρότητας μπορεί να υπολογιστεί μόνο κατά προσέγγιση.
- Ανοχή σε σφάλματα (*Fault Tolerance*): Τα *IDS* λειτουργούν μαζί με τα *firewall* σαν ασπίδα ενός πληροφοριακού συστήματος και παίζουν σημαντικό ρόλο στην ασφάλεια του, για αυτό το λόγο θα πρέπει να είναι ανθεκτικά σε σφάλματα. Αυτό σημαίνει ότι θα πρέπει να είναι υπέρ-ανθεκτικό σε επιθέσεις το ίδιο και ειδικότερα σε επιθέσεις του τύπου *Denial of Service (DoS)*.

- Επικαιρότητα (*Timeliness*): Με την έννοια επικαιρότητα εννοούμε ότι πρέπει να γρήγορο στην ανάλυση των γεγονότων έτσι ώστε να εξάγει αποτελέσματα γρηγορά και να δώσει χρόνο στον άνθρωπο που το διαχειρίζεται λογικό χρονικό διάστημα να αντιδράσει σε ένα περιστατικό ασφάλειας και να δράσει ανάλογα αποτρέποντας την επίθεση ή περιορίζοντας τις ζημίες. Ο χρόνος από την στιγμή που θα γίνει αντιληπτή η επίθεση μέχρι να γίνουν οι πρώτες ενέργειες αποτροπής της, είναι πολύτιμος. Ο αναλυτής πρέπει μέσα σε αυτό το χρόνο να εκτιμήσει την κατάσταση και να δράσει ανάλογα ενώ η επίθεση βρίσκεται σε εξέλιξη. Για αυτό τον λόγο το *IDS* πρέπει να είναι γρήγορο και επίκαιρο.

Πέραν των παραπάνω μέτρων αποδοτικότητας είναι γενικά, επιπλέον υπάρχουν και κάποιες ποσοτικές μεταβλητές με τις οποίες μπορούμε να μετρήσουμε σε αριθμούς την αποδοτικότητα και την αποτελεσματικότητα ενός *IDS*. Αρχικά θα ονομάσουμε τις μεταβλητές πάνω στις οποίος θα βασίσουμε τα μετρά αξιολόγησης ενός *IDS* [5][31] .

True positive (*TP*): Μια επίθεση ανιχνεύεται και κατηγοριοποιείται σαν επίθεση. Η τιμή των *TP* είναι γνωστή και ως *true positive rate-TPR* ρυθμός ανίχνευσης (*detection rate*) ή ευαισθησία (*sensivity*).

False positive (*FP*): Η λανθασμένη ανίχνευση μιας κανονικής δραστηριότητας ως εισβολή. Η τιμή των *FP* είναι γνωστή ως *"false positive rate-FPR"* ή λανθασμένος συναγερμός (*false alarm rate*).

True negative (*TN*): Η σωστή κατηγοριοποίηση μιας δραστηριότητας ως φυσιολογική. Η τιμή των *TN* αναφέρεται σαν "*true negative rate-TNR*"ή σαν ειδικότητα (*specificity*).

False Negative (*FN*): Η λανθασμένη κατηγοριοποίηση μιας εισβολής ως κανονική δραστηριότητα. Η τιμή των *FN* αναφέρεται και ως *false negative rate-FNR*.

Παρακάτω παρουσιάζουμε τους τύπους υπολογισμού των τιμών.

- $True\ positive\ rate\ (TPR) = \frac{TP}{TP+FN}$
- $False\ positive\ rate\ (FPR) = \frac{FP}{TN+FP} = 1 - specifity$
- $True\ negative\ rate(TNR) = \frac{TN}{TN+FP}$
- $False\ negative\ rate(FNR) = \frac{FN}{TP+FN} = 1 - sensivity$
- $Accurancy = \frac{TP+TN}{TP+FP+TN+FN}$
- $Precision\ = \frac{TP}{TP+FP}$

## 2.7 Σύνοψη

Συνοψίζοντας αυτό το κεφάλαιο μπορούμε ευκολά να συμπεράνουμε ότι ένα *IDS* είναι ένα σύστημα που έχει να αντιμετωπίσει πληθώρα απειλών και αυτό πρέπει να γίνεται γρήγορα αποτελεσματικά και χωρίς λάθη. Επίσης λόγω του ότι βρίσκεται στην πρώτη γραμμή άμυνας ενός πληροφοριακού συστήματος πρέπει να είναι ανθεκτικό σε σφάλματα, ενώ παράλληλα πρέπει σε πραγματικό χρόνο να παρακολουθεί, να αναλύει και να ελέγχει πληροφορίες από πολλές πηγές ταυτόχρονα. Αυτό οδήγησε την ερευνητική κοινότητα να στραφεί σε τεχνικές εξόρυξης δεδομένων (D*ata Mining*), τις οποίες θα αξιοποιεί το *IDS* έτσι ώστε να μπορεί να εξάγει την χρήσιμη πληροφορία μέσα από τον τεράστιο όγκο των δεδομένων που επεξεργάζεται. Αυτή η πληροφορία θα αξιοποιηθεί για την μετέπειτα ανίχνευση εισβολών και

επιθέσεων. Κάποιες από τις τεχνικές εξόρυξης δεδομένων που χρησιμοποιούνται στα *IDS* θα αναφέρουμε στην επόμενο κεφάλαιο.

# Κεφάλαιο 3ο

## Εξόρυξη Δεδομένων (*Data Mining*)

Όπως είναι γνωστό οι υπολογιστές και το διαδίκτυο έχουν γίνει κομμάτι της καθημερινότητας όλων των ανθρώπων. Είτε για δουλειά είτε για ψυχαγωγία οι υπολογιστές, οι έξυπνες συσκευές (*smartphones, tablets, smart TVs* κ.λ.π) και το διαδίκτυο είναι πλέον αναπόσπαστα κομμάτια κάθε πτυχής της ζωής μας. Όλες αυτές οι συσκευές που διασυνδέονται με τον διαδίκτυο, ο αριθμός των χρηστών που αυξάνεται χρόνο με το χρόνο και η πληθώρα των κοινωνικών και εμπορικών ιστοσελίδων έχουν εκτοξεύσει τον όγκο των δεδομένων που διακινούνται καθημερινά μέσου του διαδικτύου. Η εξάπλωση του διαδικτύου και η αύξηση του όγκου δεδομένων που διακινούνται μέσω αυτού δημιούργησε την ανάγκη για περισσότερο χώρο αποθήκευσης αυτών των δεδομένων. Όλα αυτά τα δεδομένα αποθηκεύονται σε βάσεις δεδομένων οι οποίες πλέον έχουν γιγαντιαίο όγκο και αυξάνονται συνεχώς. Όσο τα δεδομένα αυξάνονται γίνεται πιο δύσκολη η επεξεργασία τους και εξαγωγή πληροφοριών από αυτά. Η μείωση του κόστους συλλογής και επεξεργασίας δεδομένων αλλά και η επιτάχυνση αυτών των διεργασιών, συντέλεσε στην ανάπτυξη του τομέα της εξόρυξης δεδομένων. Σκοπός του κλάδου της εξόρυξης δεδομένων (*Data Mining*) είναι να μελετήσει και να αναπτύξει τεχνικές οι οποίες εξάγουν από μια μεγάλη συλλογή δεδομένων πληροφορίες και από αυτές τις πληροφορίες να αποκτήσουμε κάποια γνώση.

### 3.1 Ανακάλυψη Γνώσης από Βάσεις Δεδομένων (*Knowledge Discovery in Database*)

Όπως προαναφέραμε σκοπός της ανάλυσης των δεδομένων είναι να μπορέσουμε να αποκτήσουμε κάποια γνώση από αυτά. Ο όρος Ανακάλυψη Γνώσης από Βάσεις Δεδομένων (*Knowledge Discovery in Database -KDD*) είναι γενικότερος από τον όρο Εξόρυξη Δεδομένων (*Data Mining*). Συγκεκριμένα ο όρος *KDD* αναφέρεται στην συνολική διαδικασία που ακολουθούμε για την ανακάλυψη γνώσης μέσα από την ανάλυση δεδομένων. Πιο συγκεκριμένα αναφέρεται σε όλη την διαδικασία από την συλλογή δεδομένων μέχρι την αξιοποίηση των αποτελεσμάτων [32]. Το *Data Mining* είναι ένα από τα βήματα της *KDD* διαδικασίας τα οποία είναι τα εξής [33]:

1. Καθαρισμός Δεδομένων (*Data Cleaning*): Αφαίρεση θορύβου από τα δεδομένα αν υπάρχει.
2. Ενσωμάτωση των Δεδομένων (*Data Integration*): Όταν υπάρχουν πολλαπλές πηγές. Για παράδειγμα ένα *IDS* συγκεντρώνει δεδομένα από κόμβους δικτύου, τους υπολογιστές των χρηστών τους *server* τους πληροφοριακού συστήματος κλπ.
3. Συλλογή Δεδομένων (*Data Selection*): Συλλογή των κατάλληλων δεδομένων σχετικά με το τι θέλουμε να αναλύσουμε. Αν για παράδειγμα θέλουμε να αναλύσουμε διαδικτυακή κίνηση τότε θα πρέπει να αναλύσουμε δεδομένα κίνησης δικτύου (pcap αρχεία για παράδειγμα).
4. Μετασχηματισμός Δεδομένων (*Data Transformation*): Είναι η διαδικασία μετατροπής των δεδομένων κάτω από ένα κοινό πλαίσιο έτσι ώστε να γίνει πιο εύκολη η εξόρυξη και η επεξεργασία.
5. Εξόρυξη Δεδομένων (*Data Mining*): Η διαδικασία στην οποία εφαρμόζουμε μεθόδους έτσι ώστε να εξάγουμε συγκεκριμένα μοτίβα δεδομένων.

6. Αξιολόγηση Μοτίβου (*Pattern evaluation*): Η διαδικασία αναγνώρισης των πραγματικά χρήσιμων μοτίβων από τα οποία θα εξαχθεί η γνώση.
7. Παρουσίαση της Γνώσης (*Knowledge presentation*): Είναι το τελικό στάδιο όπου τεχνικές αναπαράστασης και οπτικοποίησης εφαρμόζονται για να παρουσιαστεί και να είναι κατανοητή η γνώση που εξήχθη στους χρήστες.

Τα βήματα ένα έως τέσσερα είναι διαφορετικές μορφές προ-επεξεργασίας και έπειτα εφαρμόζονται οι τεχνικές εξόρυξης.

Τέλος, σε πολλούς τομείς όπως η βιομηχανία και η επιστημονική κοινότητα χρησιμοποιούν τον όρο *Data Mining* για να περιγράψουν όλη την διαδικασία του *KDD* οπότε γενικά τα *Data Mining* μπορεί να περιγράφει σαν μια διαδικασία ανακάλυψης μοτίβων από μεγάλους όγκους δεδομένων.

## 3.2 Ποια είδη δεδομένων μπορούμε να εξορύξουμε;

Η εξόρυξη δεδομένων (*Data Mining*) είναι μια μεθοδολογία η οποία αναπτύχθηκε με σκοπό να μπορεί να εφαρμοστεί σε οποιοδήποτε τύπο δεδομένων. Ωστόσο οι αλγόριθμοι και οι τεχνικές προσέγγισης που χρησιμοποιούνται συνήθως διαφέρουν για τους διαφόρους τύπους δεδομένων που χρησιμοποιούνται. Κάποια από τα είδη των δεδομένων στα οποία μπορούν να εφαρμόσουμε τεχνικές *Data Mining* είναι οι ροές δεδομένων, σε δεδομένα μορφής κειμένου (*text*) , διαδικτυακής κίνησης, δεδομένα ακολουθίας (π.χ. ιστορικά γεγονότα, χρηματοοικονομικές συναλλαγές), χωρικά δεδομένα (π.χ. χάρτες), δεδομένα μέσων (π.χ. εικόνα, ήχο) και δεδομένα παγκοσμίου ιστού. Όλα τα παραπάνω είδη δεδομένων έχουν και τις αντίστοιχες βάσεις δεδομένων όπως αναφέρεται στο [34]. Συνοπτικά αυτές είναι οι κάτωθι:

- Flat files: Αυτού του είδους βάσεις δεδομένων είναι απλοί φάκελοι σε μορφή κειμένου (*text*).
- Relational Databases: Είναι μια συλλογή από πίνακες, και κάθε πίνακας έχει ένα όνομα. Κάθε πίνακας έχει γραμμές και στήλες. Οι στήλες περιέχουν κάποιο χαρακτηριστικό (π.χ. μια στήλη περιέχει το όνομα και μια δεύτερη το επίθετο πελάτη) και οι γραμμές το πλήθος των στοιχείων που υπόκεινται σε αυτό το χαρακτηριστικό (δηλαδή τα ονόματα όλων των πελάτων).
- Data Warehouses: Είναι μια μεγάλη αποθήκη δεδομένων που συλλεγεί δεδομένα από διαφορετικές πηγές
- Transaction Databases: Αυτή η βάσει δεδομένα αποτελείται από εάν σύνολο καταγραφών οι οποίες αναπαριστούν συναλλαγές (*transactions*). Κάθε συναλλαγή έχει μια χρονοσφραγίδα, ένα μοναδικό χαρακτηριστικό και ένα σύνολο στοιχείων. Ένα τέτοιο είδος βάσεις δεδομένων είναι και τα καλάθια αγορών που υπάρχουν στους διαδικτυακούς ιστότοπους πώλησης αγαθών. Όταν για παράδειγμα γίνει μια αγορά στην βάσει αποθηκεύεται η ώρα που έγινε η συναλλαγή (χρονοσφραγίδα), ο αριθμός της παραγγελίας που είναι μοναδικός για κάθε πελάτη και τα αντικείμενα που αγόρασε.
- Multimedia Databases: Βάσεις δεδομένων οι οποίες είναι μεγάλες σε όγκο διότι σε αυτές αποθηκεύονται δεδομένα εικόνας, ήχου, βίντεο και κειμένου.
- Time Series Databases: Σε αυτές τις βάσεις συνήθως αποθηκεύονται δεδομένα για τα οποία μεγάλη σημασία παίζει η χρονική σειρά. Τέτοιες βάσεις δεδομένων χρησιμοποιούνται από τα πληροφοριακά σύστημα για την καταγραφή των *log*. Επίσης στο χρηματιστήριο χρησιμοποιούνται τέτοιου είδους βάσεις για την καταγραφή της πορείας της τιμής των μετοχών σχετικά με τον χρόνο.
- Spatial Databases: Βάσεις δεδομένων οι οποίες αποθηκεύουν γεωγραφικά δεδομένα και χρησιμοποιούνται για κυρίως για χάρτες.

- World Wide Web: Οι βάσεις δεδομένων στον παγκόσμιο ιστό οργανώνονται σε αλληλοσυνδεόμενα αρχεία τα οποία περιέχουν κάθε μορφής δεδομένα όπως εικόνα, βίντεο, ήχο, κείμενο.

## 3.3 Είδη Αντικειμένων και Χαρακτηριστικών

Στην προηγουμένη ενότητα μιλήσαμε για βάσεις δεδομένων και για δεδομένα τα οποία αυτές περιέχουν. Πριν επεκταθούμε σε περεταίρω έννοιες όπως επιλογή χαρακτηριστικών (*feature selection*) καλό είναι να αναφερθούμε στο τι ορίζουμε ως αντικείμενο και τι ως χαρακτηριστικό. Μια συλλογή δεδομένων απαρτίζεται από αντικείμενα δεδομένων (*data objects*). Τα αντικείμενα αυτά αναπαριστούν οντότητες. Σε μια βάσει δεδομένων ενός σχολείου για παράδειγμα οντότητες θεωρούνται οι μαθητές, οι δάσκαλοι και τα μαθήματα που διδάσκονται. Κάθε οντότητα περιγράφεται από κάποια χαρακτηριστικά (π.χ. ηλικία φίλο). Όλα αυτά μπαίνουνε σε ένα πίνακα. Σε αυτόν τον πίνακα οι οντότητες είναι οι γραμμές του πίνακα και τα χαρακτηριστικά είναι οι στήλες.

Τα χαρακτηριστικά περιγράφουν κάποιες ιδιότητες ενός αντικειμένου κάτι δηλαδή που χαρακτηρίζει το αντικείμενο. Αν για παράδειγμα στον πίνακα με τους μαθητές έχουμε την στήλη «διαγωγή μαθητή», τότε αυτό το χαρακτηριστικό περιγράφει την διαγωγή του μαθητή που βρίσκεται σε κάθε γραμμή του πίνακα. Στην διεθνή βιβλιογραφία συναντούμε τον όρο χαρακτηριστικό ως *attribute, variable, feature* [33]. Στον κλάδο του *Data Mining* οι επιστήμονες αρέσκονται στο να χρησιμοποιούν τον όρο *attribute*, στον κλάδο της στατιστικής αναφέρεται ως *variable* ενώ στο κλάδο της μηχανικής μάθησης ως *feature*. Εμείς εδώ θα το αναφέρουμε κατά σύμβαση ως *feature.*

Κάθε χαρακτηριστικό μπορεί να έχει διακεκριμένες , συνεχής (ή άπειρες) ή σύνθετες τιμές. Τα συνεχή χαρακτηριστικά (*continuous feature*) μπορεί έχουν τιμές από όλο των χώρο των πραγματικών αριθμών πράγμα που σημαίνει ότι ο αριθμός των πιθανών τιμών τους να είναι άπειρος. Τα σύνθετα χαρακτηριστικά (*complex features*) μπορεί να είναι δομές σύνθετων αριθμών (π.χ. χ+yi). Τα χαρακτηριστικά με διακεκριμένες τιμές (*discrete feature*) έχουν περιορισμένο αριθμό τιμών. Μπορούν να χωριστούν επίσης σε ονομαστικά (*nominal*) και τακτικά (*ordinal*) [35]. Τέλος έχουμε τα δυαδικά (*binary*) και τα αριθμητικά (*numerical*) [36].

- Ονομαστικά χαρακτηριστικά (*Nominal features*): Οι τιμές των χαρακτηριστικών αυτών είναι σύμβολα ή ονόματα πραγμάτων. Κάθε τιμή αναπαριστά μια κατηγορία, κωδικό ή κατάσταση.
- Δυαδικά χαρακτηριστικά (*Binary features*): τα οποία είναι υποκατηγορία των ονομαστικών και περιγράφουν δυο καταστάσεις οι οποίες είναι οι 0 (*false*) και 1 (*true*). Το 0 δηλώνει ότι το χαρακτηριστικό δεν υπάρχει και το 1 ότι υπάρχει.
- Τακτικά χαρακτηριστικά (*Ordinal features*): Οι τιμές τους έχουν λογική σειρά ή υποδηλώνουν κατάταξη μεταξύ τους. Κλασσικό παράδειγμα τέτοιου είδους χαρακτηριστικών η τιμή που βάζουμε στον πίνακα κατάταξης ανάλογα με την θέση που τερμάτισε (π.χ. πρώτος τιμή 1 , δεύτερος τιμή 2 κ.λ.π).
- Αριθμητικά χαρακτηριστικά (*Numerical attributes*): Αναπαριστούν τιμές ή ακεραίους και είναι μετρήσιμες ποσότητες. Επίσης είναι ποσοτικά χαρακτηριστικά διότι μετρούν την ποσότητα κάθε αντικειμένου του πίνακα.

## 3.4 Μέθοδοι Εξόρυξης Δεδομένων

Στην εξόρυξη δεδομένων υπάρχουν διάφορες μέθοδοι που εφαρμόζονται για την πρόβλεψη και την περιγραφή των μοτίβων δεδομένων από τα οποία θέλουμε να εξαγάγουμε πληροφορία [32][37]. Παρακάτω θα περιγράψουμε αυτές τις μεθόδους περιληπτικά.

- Μέθοδοι Παλινδρόμησης (*Regression Methods*): Είναι μια στατιστική μέθοδος η οποία συνήθως χρησιμοποιείται για αριθμητικές προβλέψεις. Η εφαρμογή αυτών των μοντέλων έχει πολλές εφαρμογές στον πραγματικό κόσμο. Για παράδειγμα αυτήν η μέθοδος εφαρμόζεται για την πρόβλεψη της τάσης της αγοράς και την ζήτηση ενός προϊόντος από του καταναλωτές [38][39][40].
- Μέθοδοι Σύνοψης (*Summarization Methods*): Σε αυτήν την μέθοδο γίνονται ενέργειες έτσι ώστε ένα μεγάλο σύνολο δεδομένων να μπορεί να περιγράφει από κάποια χαρακτηριστικά και ένα μικρότερο υποσύνολο δεδομένων. Αυτές οι μέθοδοι εφαρμόζονται σε διαδραστικές έρευνες ανάλυσης δεδομένων και για την παραγωγή αυτοματοποιημένων αναφορών [41][42].
- Σχεσιακές Μέθοδοι (*Association Methods*): Αυτές οι μέθοδοι προσπαθούν να βρουν σχέσεις εξάρτησης μεταξύ των διαφόρων μεταβλητών. Αυτά τα μοντέλα συχνά αναφέρονται και ως Μέθοδοι Συνδέσμων (*Link Methods*) ή Μέθοδοι Εξάρτησης (*Dependency Methods*) [43][44].
- Μέθοδοι Ταξινόμησης (*Classification Methods*): Με τις μεθόδους ταξινόμησης τα δεδομένα κατηγοριοποιούνται σε κλάσεις. Με απλά λόγια τα δεδομένα ανάλογα με κάποια χαρακτηριστικά που έχουν τοποθετούνται σε κατηγορίες για ευκολότερη ανάλυση και επεξεργασία [**Error! Reference source not found.**].
- Μέθοδοι Συσταδοποίησης (*Clustering Methods*): Οι μέθοδοι συσταδοποίησης ταξινομούν τα δεδομένα σε ομάδες με σκοπό τα δεδομένα που βρίσκονται μέσα στην ομάδα να έχουν μεγάλη ομοιότητα (*maximum intraclass similarity*) μεταξύ τους αλλά τα δεδομένα της μιας ομάδας με κάποιας άλλης να ξεχωρίζουν διακριτά (*minimum interclass similarity*) [46][**Error! Reference source not found.**].
- Μέθοδοι ακραίων τιμών (*Outliers Methods*): Ένα σύνολο δεδομένων μπορεί να περιέχει αντικείμενα τα οποία δεν ταιριάζουν με τα υπόλοιπα δεδομένα ή την γενική συμπεριφορά του μοντέλου. Αυτά τα δεδομένα τα αποκαλούμε ακραίες τιμές (*outliers*). Για αυτό τον λόγο έχουν αναπτυχθεί τεχνικές για την εύρεση αυτών των ακραίων τιμών. Κάποιες φορές οι ακραίες τιμές έχουν περισσότερη σημασία από τα υπόλοιπα δεδομένα του συνόλου. Οι ακραίες τιμές μπορούν να ανιχνευθούν χρησιμοποιώντας στατιστικά μοντέλα ή μοντέλα μέτρησης απόστασης (αν κάποιο στοιχείο ενός συνόλου απέχει αρκετά από τα υπόλοιπα στοιχεία τότε θεωρείται ακραία τιμή) [48][49].

## 3.5 Τεχνολογίες Εξόρυξης Δεδομένων

Για την εξαγωγή πληροφορίας και απόκτηση γνώσης ο κλάδος της εξόρυξης δεδομένων υιοθέτησε τεχνολογίες που εφαρμοστήκαν σε τομείς όπως η στατιστική, οι βάσεις δεδομένων και η μηχανική μάθηση [32][33][50].

- Στατιστική (*Statistic*): Η στατιστική ανάλυση δεδομένων είναι μια από τις πιο καλά εδραιωμένες και μελετημένες τεχνικές για εξόρυξη δεδομένων και προσφέρει μια ποικιλία μεθόδων για τον σκοπό αυτό. Στην στατιστική έχουν αναπτυχθεί εργαλεία τα οποία βασίζονται σε δεδομένα και στατιστικά μοντέλα χρησιμοποιούνται για να προβλέψουν μια πιθανή μελλοντική κατάσταση (π.χ. πρόβλεψη τιμών μετοχών). Επίσης οι στατιστικές μέθοδοι χρησιμοποιούνται για την συνοχή

ή περιγραφή συλλογών δεδομένων αλλά και για την επικύρωση αποτελεσμάτων που προήλθαν από την εξόρυξη δεδομένων. Το μεγάλο μειονέκτημα κάποιων από αυτές τις μεθόδους είναι ότι σε περίπτωση εφαρμογής τους σε μεγάλο αριθμό δεδομένων όπως και συνήθως συμβαίνει στην εξόρυξη δεδομένων, τότε αυξάνεται κατά πολύ το υπολογιστικό κόστος. Για τον λόγο χρειάζεται πολύ προσοχή και μελέτη κατά τον σχεδιασμό και την υλοποίηση των αλγορίθμων στατιστικής ανάλυσης [51].

- Βάσεις Δεδομένων (*Databases*): Ο κλάδος της επιστήμης που ασχολείται με τα συστήματα βάσεων δεδομένων έχει εστιάσει την προσοχή του στην δημιουργία, συντήρηση και χρήση βάσεων δεδομένων για οργανισμούς και τελικούς χρήστες. Τα συστήματα, τα μοντέλα και οι γλώσσες (*query languages*) που έχουν αναπτυχθεί για χρήση σε βάσεις δεδομένων έχουν την δυνατότητα επεξεργασίας μεγάλου όγκου δεδομένων. Όλα τα παραπάνω μπορούμε να τα εκμεταλλευτούμε και για εξόρυξη δεδομένων λόγω της αποτελεσματικότητας και της επεκτασιμότητας που προσφέρουν [52].
- Μηχανική Μάθηση (*Machine Learning*): Ο κλάδος της μηχανικής μάθησης ασχολείται με το πως μπορούν οι υπολογιστές να μάθουν πράγματα από ένα σύνολο δεδομένων που τους δίνεται. Η επιστημονική έρευνα αυτού του κλάδου έχει επικεντρωθεί στην ανάπτυξη προγραμμάτων που κάνουν χρήση αλγορίθμων μηχανικής μάθησης. Με αυτά τα προγράμματα οι υπολογιστές θα εκπαιδεύονται και θα μαθαίνουν από τα δεδομένα που έχουν έρχονται σαν είσοδο στο σύστημα και έπειτα θα μπορούν να παίρνουν αποφάσεις. Όλη η παραπάνω διαδικασία πρέπει να είναι αυτοματοποιημένη και αν χρειάζεται ανθρώπινη παρέμβαση αυτήν θα πρέπει να είναι σε μικρό βαθμό [53]. Η μηχανική μάθηση θα παίξει σημαντικό ρόλο σε πολλές πτυχές του κλάδου της πληροφορικής, στον κλάδο της εξόρυξης δεδομένων. Επιπλέον θα παίξει σημαντικό ρόλο στα σύστημα ασφάλειας πληροφοριακών συστημάτων όπου πληθώρα νέων απειλών εμφανίζονται καθημερινά και πλέον θα πρέπει τα συστήματα να αρχίζουν να αντιλαμβάνονται από μόνα τους τι είναι κακό και τι καλό. Τέτοια συστήματα έχουν αρχίσει να αναπτύσσονται και να δοκιμάζονται διότι είναι επιτακτική ανάγκη. Οι υπάρχουσες μέθοδοι αποτυγχάνουν να προσφέρουν ικανοποιητικό βαθμό προστασίας και εξαρτώνται κυρίως από τον ανθρώπινο παράγοντα (π.χ. *signature based antivirus systems*). Η μηχανική μάθηση και οι τεχνικές που έχουν αναπτυχθεί πάνω σε αυτήν την φιλοσοφία θα περιγράφουν σε επόμενο κεφάλαιο.

## 3.6 Προβλήματα που αντιμετωπίζουμε στην εξόρυξη δεδομένων

Είναι πολλοί οι τομείς στους οποίους μπορούμε να χρησιμοποιήσουμε την εξόρυξη δεδομένων για την εξαγωγή πληροφοριών και να αποκτήσουμε γνώση για κάτι που χρειαζόμαστε. Αυτό έχει ως συνέπεια ότι από τομέα σε τομέα τα δεδομένα διαφέρουν μεταξύ τους. Για παράδειγμα είναι διαφορετικό να συλλέγεις δεδομένα από μέσα κοινωνικής δικτύωσης στο διαδίκτυο και διαφορετικό να συλλέγεις δεδομένα *logs* από υπολογιστές ή διαδικτυακή κίνηση. Για αυτό τον λόγο πρέπει οι σχεδιαστές εργαλείων εξόρυξης δεδομένων να μελετήσουν καλά τις πηγές και το είδος των δεδομένων που αυτές παράγουν έτσι ώστε να καταλήξουν στην κατάλληλη μεθοδολογία και αλγορίθμους που θα υλοποιήσουν [33][34][35][36][37][54]. Σε περίπτωση εξόρυξης δεδομένων από πολλές διαφορετικές πηγές τότε έχουμε πολλών διαφορετικών ειδών δεδομένα και το εργαλείο εξόρυξης δεδομένων που θα αναπτυχθεί πρέπει να εφαρμόζει μεθοδολογίες οι οποίες να καλύπτουν όλους τους διαφορετικούς τύπους δεδομένων. Αυτό θα έχει ως συνέπεια ότι για κάποιους τύπους δεδομένων θα έχουμε καλά αποτελέσματα ενώ για κάποιους άλλους όχι, διότι δεν υπάρχει μεθοδολογία ή αλγόριθμος που να είναι

καλός για όλα, μέχρι στιγμής τουλάχιστον. Τα δεδομένα που μπορούμε να συλλέξουμε είναι πολλά και πολλές φορές περισσότερα από ότι μπορεί να χρειαζόμαστε.

Η φιλοσοφία που επικρατεί στις μέρες είναι να συλλέγουμε όσα πιο πολλά δεδομένα μπορούμε και όταν τα χρειαστούμε να τα επεξεργαστούμε και να τα αναλύσουμε. Εδώ όμως τίθεται ένα άλλο θέμα το οποίο είναι το κατά πόσο μπορούμε να επεξεργαστούμε αποτελεσματικά τόσο μεγάλο όγκο δεδομένων σε λογικό χρόνο. Πολλοί αλγόριθμοι έχουν αναπτυχθεί και βελτιστοποιηθεί για να είναι πιο γρήγοροι αλλά όταν έχουμε τεράστιο όγκο δεδομένων. Στην πλειοψηφία τους αυτοί οι αλγόριθμοι δεν μπορούν να χειριστούν αποτελεσματικά μεγάλους όγκους δεδομένων και είναι αργοί [55]. Μέθοδοι δειγματοληψίας έρχονται να δώσουν εν μέρη λύση σε αυτά τα προβλήματα χρησιμοποιώντας μια αντιπροσωπευτική ποσότητα από ένα μεγαλύτερο σύνολο δεδομένων. Και εδώ όμως εγείρονται ερωτήματα το κατά πόσο είναι τα δείγματα αντιπροσωπευτικά του γενικότερου συνόλου. Παρόλα αυτά η εξόρυξη δεδομένων δεν παύει να είναι μια υπολογιστικά και χρονικά κοστοβόρα διαδικασία.

Ένα άλλο σημαντικό κομμάτι της εξόρυξης δεδομένων είναι ανάδραση που δίνουν στον χρήστη τα εργαλεία που χρησιμοποιεί. Η ανάλυση που θα γίνει από το εκάστοτε εργαλείο θα πρέπει να είναι ευκολά κατανοητή από τον χρήστη. Τεχνικές οπτικοποίησης των αποτελεσμάτων όπως γραφικές παραστάσεις και διαγράμματα διευκολύνουν την γρηγορότερη και εις βάθος κατανόηση του τι έχει αναλυθεί [50][56][57]. Επίσης βοηθούν την γρηγορότερη εξαγωγή συμπερασμάτων η οποία με την σειρά της οδηγεί στην γρηγορότερη λήψη αποφάσεων.

Τέλος υπάρχουν ηθικά ζητήματα και ζητήματα ασφάλειας. Η συλλογή δεδομένων από διαφορετικές πηγές πολλές φορές οδηγεί σε συγκλονιστικά συμπεράσματα τα οποία δεν θα μπορούσαν να είχαν επιτευχθεί αλλιώς. Οι ανησυχίες που υπάρχουν γενικότερα στον κόσμο είναι το ποσά δεδομένα πρέπει να συλλέγονται, τι δεδομένα συλλέγονται, αν θα πρέπει να συλλέγονται και από ποιους συλλέγονται. Ιδιαίτερες ανησυχίες υπάρχουν για την συλλογή προσωπικών και ευαίσθητων δεδομένων (π.χ. αριθμός διαβατηρίου, ιατρικός φάκελος). Ειδικά δεδομένα όπως τα τελευταία θα πρέπει να διασφαλίζονται με τον καλύτερο δυνατό τρόπο και να γίνεται προσεκτική χρήση αυτών. Αλλιώς υπάρχει κίνδυνος τα άτομα στα οποία ανήκουν αυτά τα δεδομένα να εκτεθούν ανεπανόρθωτα [58][59][60][61].

## 3.7 Εξόρυξη Δεδομένων και Συστήματα Ανίχνευσης Εισβολών

Όπως αναφέραμε και προηγουμένως η εξόρυξη δεδομένων έχει εφαρμογή σε πολλούς τομείς. Ένας από αυτούς είναι και τομέας της ασφάλειας πληροφοριακών συστημάτων και πιο συγκεκριμένα στα συστήματα ανίχνευσης εισβολών (*IDS*). Στο προηγούμενο κεφάλαιο αναφέραμε κάποιες τεχνικές που χρησιμοποιούν τα *IDS* και τα προβλήματα που αντιμετωπίζουν. Ο μεγάλος όγκος πληροφοριών που καλείται να αναλύσει ένα *IDS* ώθησε την ερευνητική κοινότητα να στραφεί σε αξιοποίηση τεχνικών που χρησιμοποιούνται στην εξόρυξη δεδομένων για να επιλύσει τα προβλήματα αυτά. Αρκετές έρευνες έχουν γίνει προς αυτήν την κατεύθυνση στο [62] οι συγγραφείς κάνουν μια αναφορά στις τεχνικές ανίχνευσης ανωμαλιών. Εκεί τονίζουν ότι η εξόρυξη δεδομένων μπορεί να βελτιώσει την διαδικασία ανίχνευσης εισβολών επικεντρώνοντας την προσοχή στην ανίχνευση ανωμαλιών. Επίσης η αναγνώριση και ο διαχωρισμός των «συνόρων» μεταξύ της φυσιολογικής και μη φυσιολογικής διαδικτυακής κίνησης, θα βοηθήσει τον αναλυτή του διαδικτύου να ξεχωρίσει ευκολότερα μια διαδικτυακή κίνηση η οποία περιέχει μια επίθεση από μια καθημερινή κανονική διαδικτυακή κίνηση. Στο [6] τονίζεται η σημασία και η συνεισφορά των μεθόδων εξόρυξης δεδομένων για ένα *IDS* και τα πλεονεκτήματα που αυτή φέρει ειδικά στην εφαρμογή της σε μεθόδους ταξινόμησης (*classification*). Έπειτα έχουν προταθεί και διάφορες δομές (*frameworks*) πάνω στις οποίες μπορεί να δημιουργηθεί ένα *IDS* το οποίο χρησιμοποιεί τεχνικές

εξόρυξης δεδομένων. Στο [63] προτείνεται από τους συγγραφείς μια δομή ή οποία χρησιμοποιεί αλγορίθμους εξόρυξης δεδομένων έτσι ώστε να αναγνωρίσει τα μοτίβα δραστηριοτήτων που εμφανίζονται συχνά. Έπειτα αυτά τα μοτίβα χρησιμοποιούνται για να γίνει επιλογή χαρακτηριστικών (*features*) τα οποία θα χρησιμοποιηθούν από τους αλγόριθμους ταξινόμησης (*classifiers*). Οι αλγόριθμοι ταξινόμησης με την σειρά τους θα χρησιμοποιούμουν για να χτίσουν τα μοντέλα ανίχνευσης εισβολών. Παρόμοια με το προηγούμενο, άλλη μια δομή προτείνεται στο [64]. Εδώ η κεντρική ιδέα είναι να δημιουργήσουν προγράμματα εξόρυξης δεδομένων τα οποία θα μαθαίνουν κανόνες οι οποίοι θα περιγράφουν με ακρίβεια την ομαλή και την μη ομαλή συμπεριφορά του δικτύου. Η δομή τους περιέχει προγράμματα για δημιουργία ταξινομητών (*classifiers*) και μετά-ταξινομητών (*meta classifiers*), σχεσιακών κανόνων (association rules), για ανάλυση συνδέσμων (*link analysis*) , την συχνότητα συμβάντων (*frequent analysis*), την ανάλυση αλληλουχίας (*sequence analysis*) και ένα περιβάλλον υποστήριξης γενικότερα το οποίο ενσωματώνει δομικά στοιχεία τα οποία βοηθούν στην δημιουργία και αξιολόγηση μοντέλων ανίχνευσης. Μια ακόμα πρωτοπόρα δομή, το MADAM ID, παρουσιάζεται στο [28]. Αυτή η δομή χρησιμοποιεί αλγορίθμους εξόρυξης δεδομένων για να υπολογίσει τα μοτίβα δραστηριοτήτων και να εξάγει χαρακτηριστικά από αυτά τα μοτίβα. Έπειτα εφαρμόζει αλγορίθμους μηχανικής μάθησης για να δημιουργήσει κανόνες ανίχνευσης εισβολών. Τα αποτελέσματα από τα πειράματα που έγιναν έδειξαν ότι τα δεδομένα που εξορύχθηκαν μπορούν να είναι αξιόπιστα και μπορούν να χρησιμοποιηθούν για την δημιουργία μοντέλων ανίχνευσης ανωμαλιών.

Μελέτες για την εφαρμογή τεχνικών εξόρυξη δεδομένων έχουν γίνει και για τα σύστημα ανίχνευσης εισβολών δικτύου (*Network IDS-NIDS*). Στο [65] αναφέρονται κάποιες μέθοδοι εξόρυξης δεδομένων που εφαρμόζονται για *NIDS* και προτείνουν μια προσέγγιση η οποία πιστεύουν ότι θα βοηθήσει στην δημιουργία καλυτέρων και αποδοτικότερων *IDS*. Το [66] περιγράφει ένα σύστημα το οποίο είναι ικανό να ανιχνεύσει διαδικτυακές εισβολές χρησιμοποιώντας τεχνικές συσταδοποίησης (*clustering*). Το συγκεκριμένο σύστημα εφαρμόζει την τεχνική της συσταδοποίησης χωρίς επίβλεψη (*unsupervised*) για να ομαδοποιήσει συμπεριφορές διαδικτυακής κίνησης, να ανιχνεύσει διαφορετικές συμπεριφορές από τις συνηθισμένες και να τις ομαδοποιήσει σαν ακραίες τιμές (*outliers*).

Ερευνά έχει γίνει και για την ανάπτυξη *IDS* που θα μπορούν να λειτουργούν σε περιβάλλοντα πραγματικού χρόνου. Στο [67] γίνεται μια αναφορά σε μεθόδους εξόρυξης δεδομένων για χρήση σε συστήματα ανίχνευσης εισβολών που θα ανταποκρίνονται σε πραγματικό χρόνο. Επίσης υλοποιήθηκε ένα πρότυπο σύστημα το οποίο έχει την δυνατότητα να ανιχνεύει εισβολές σε επίπεδο δικτύου ή συστήματος. Τέλος στο [68] γίνεται συνδυασμός τεχνικών εξόρυξης δεδομένων και μηχανικής μάθησης που έχει σαν αποτέλεσμα την μείωση των λανθασμένων συναγερμών (*false positives*) και γενικότερα να βελτιώσουν την ποιότητα των συναγερμών (*alerts*) σε ένα *IDS*.

## 3.8 Σύνοψη

Όπως είδαμε η σύγχρονη εποχή των υπολογιστών και της πληροφορικής γενικότερα, γίνεται διακίνηση μεγάλου όγκου δεδομένων. Αυτός ο όγκος δεδομένων δεν αρκεί μόνο να αποθηκευτεί κάπου αλλά πρέπει και να αναλυθεί άμεσα ή έμμεσα. Ο κλάδος της εξόρυξης δεδομένων έρχεται να δώσει λύσεις στον τρόπο επεξεργασίας και παρουσίασης αυτών των δεδομένων με ποικιλία τεχνικών και μεθόδων. Πολλές από αυτές τις τεχνικές εστιάζουν στην όλο και λιγότερη εμπλοκή του χρήστη στην διαδικασία της ανάλυσης και προσπαθούν να αυτοματοποιήσουν το μεγαλύτερο μέρος της διαδικασίας, αφήνοντας στον αναλυτή περισσότερο χρόνο για την λήψη αποφάσεων και αντίδρασης. Όλες οι τεχνικές αυτοματοποίησης της διαδικασίας στρέφονται σε αυτό που αποκαλούμε μηχανική μάθηση (*machine learning*). Η μηχανική μάθηση σαν τεχνολογία είναι η νέα τάση σε συστήματα ανίχνευσης εισβολών. Εκεί

βασίζονται οι περισσότερες πλέον μελέτες για αυτοματοποίηση διαδικασιών και λήψη αποφάσεων οι οποίες θα πραγματοποιούνται από μηχανές. Για την μηχανική μάθηση θα αναφερθούμε εκτενέστερα στο επόμενο κεφάλαιο.

# Κεφάλαιο 4ο

## Μηχανική μάθηση

Όπως είδαμε στο προηγούμενο κεφάλαιο ο μεγάλος όγκος δεδομένων που πρέπει να επεξεργαστούμε και να αναλύσουμε είναι ένα από τα μεγαλύτερα προβλήματα της εποχής μας. Αυτό οδήγησε την επιστημονική κοινότητα στην ανάπτυξη τεχνικών αυτοματοποίησης των διαδικασιών επεξεργασίας αυτού του τεράστιου όγκου δεδομένων και ως συνέπεια ενός ολοκλήρου κλάδου της πληροφορικής και της τεχνητής νοημοσύνης. Η επιστήμη της πληροφορικής που ασχολείται με την αυτοματοποίηση διαδικασιών και λήψη αποφάσεων οι οποίες θα γίνονται από υπολογιστές ονομάζεται μηχανική μάθηση (*machine learning*). Ο *Alan Turing* ήταν ο πρώτος ο οποίος προσπάθησε να προσδιορίσει αν ένας υπολογιστής έχει νοημοσύνη δημιουργώντας το "*Turing Test*" [69]. Η πρώτη προσπάθεια δημιουργίας προγράμματος μηχανικής μάθησης έγινε το 1959 από τον *Arthur Samuel* [70] ο οποίος έφτιαξε ένα πρόγραμμα το οποίο βασιζόταν στο παιχνίδι της ντάμας (*game of checkers*). Ο υπολογιστής της IBM στον οποίο έτρεχε τότε το πρόγραμμα, βελτίωνε την γνώση του όσο περισσότερο έπαιζε το παιχνίδι, μελετώντας τις στρατηγικές που ήταν πιο αποδοτικές και κατέληγαν σε νίκες, αποθηκεύοντας αυτές στην μνήμη του. Από εκεί και πέρα υπήρξε μεγάλη αφοσίωση από την επιστημονική κοινότητα και εξελίχθηκαν πολλές μέθοδοι και αλγόριθμοι για το πως οι υπολογιστές να μπορούν να μάθουν και να πάρουν αποφάσεις. Γενικά η μηχανική μάθηση έχει ως σκοπό να καταστήσει του υπολογιστές ικανούς να μπορούν να είναι εν μέρη αυτόνομοι και να μην χρειάζεται ο επαναπαραμετροποιησή τους κατά τακτά χρονικά διαστήματα.

Πριν όμως επεκταθούμε περαιτέρω στο τεχνικό κομμάτι καλό είναι να κάνουμε μια περιγραφή του όρου «μάθηση». Για να μπορέσουν οι επιστήμονες να προχωρήσουν στην ανάπτυξη τεχνικών μηχανικής μάθησης έπρεπε να μελετήσουν καλά το ζωικό βασίλειο και κυρίως το ανθρώπινο είδος για να κατανοήσουν τις βασικές αρχές της μάθησης και πως αυτήν εξελίσσεται. Ο άνθρωπος και τα ζώα χτίζουν την γνώση τους βασιζόμενα σε εμπειρίες που έχουν . Αυτές οι εμπειρίες είναι τα δεδομένα που λαμβάνονται κάθε στιγμή και καθώς μεγαλώνει ένα ον αποκτά όλο και περισσότερες εμπειρίες οι οποίες συνεισφέρουν στην ευφυΐα του [71]. Κάθε ζώο για να αντιμετωπίσει τις διάφορες καταστάσεις που προκύπτουν χρησιμοποιεί αυτά που έχει μάθει μέχρι εκείνη την στιγμή και αντιδρά ανάλογα. Η μάθηση

είναι αυτή η οποία δίνει σε κάθε ζώο την δυνατότητα να προσαρμόζεται σε νέες καταστάσεις και να εξελίσσεται. Η μάθηση και η προσαρμογή είναι δύο θεμελιώδη στοιχεία της νοημοσύνης.

Βασιζόμενοι στα παραπάνω οι ειδικοί προσπαθούν με την μηχανική μάθηση να κάνουν τους υπολογιστές ικανούς να προσαρμόζονται και να τροποποιούν τις ενέργειες που θα κάνουν ανάλογα με την περίσταση που αντιμετωπίζουν έτσι ώστε να αυξήσουν την αποτελεσματικότητά τους. Οι μέθοδοι που έχουν αναπτυχθεί είναι πολλοί και έχουν εφαρμογή σε πολλά προβλήματα. Βέβαια καλό είναι να επισημάνουμε ότι προσοχή χρειάζεται στην επιλογή του επιλεχθέντος αλγορίθμου που θα χρησιμοποιηθεί για την επίλυση κάποιου προβλήματος. Αυτό διότι κάποιοι αλγόριθμοι μηχανικής μάθησης έχουν μεγάλη υπολογιστική πολυπλοκότητα και δεν είναι αποτελεσματικοί για επεξεργασία μεγάλου όγκου δεδομένων. Για αυτό το λόγο χρειάζεται μεγάλη προσοχή στο τι αλγόριθμο επιλέγουμε και την υπολογιστική πολυπλοκότητα αυτού [72].

## 4.1 Είδη Μηχανικής Μάθησης

Τα είδη μηχανικής μάθησης χωρίζονται σε κατηγορίες. Αυτές οι κατηγορίες προκύπτουν από τους αλγόριθμους και τον τρόπο με τον οποίο αυτοί μαθαίνουν. Υπάρχουν τέσσερις κατηγορίες μηχανικής μάθησης η επιβλεπόμενη (*supervised*), η μη επιβλεπόμενη (*unsupervised*), η ημί-επιβλεπόμενη (*semi supervised*), η ενισχυτική μάθηση (*reinforced*) και η εξελικτική μάθηση (*evolutionary*).

### 4.1.1 Επιβλεπόμενη Μάθηση (*Supervised Learning*)

Σε αυτήν την κατηγορία κατά την διαδικασία της μάθησης παρέχουμε στον αλγόριθμο σαν είσοδο ένα σύνολο δεδομένων και τις αντίστοιχες σωστές απαντήσεις-αποτελέσματα (δηλαδή την έξοδο) για κάθε ένα από αυτά τα στοιχεία εισόδου. Έτσι κατασκευάζει μια συνάρτηση η οποία αντιστοιχεί τιμές εισόδου σε τιμές εξόδου. Ο αλγόριθμος βασιζόμενος στις τιμές εισόδου και εξόδου που έχει μάθει, γενικεύει την συνάρτηση απόφασης με στόχο να προβλέπει πλέον τιμές εξόδου για κάθε είσοδο τιμής που θα δέχεται στο μέλλον. Η διαδικασία της εκπαίδευσης του αλγορίθμου πρέπει να γίνει έτσι ώστε να μπορεί ο αλγόριθμος φτάσει σε ένα επιθυμητό επίπεδο που να μπορεί να κάνει σωστές προβλέψεις με όσο το δυνατό μεγαλύτερη ακρίβεια για τυχαίες τιμές εισόδου. Στις μεθόδους επιβλεπόμενης μάθησης τα δεδομένα έχουν κάποια ετικέτα (*labeled data*) που τα προσδιορίζει. Αυτή η ετικέτα μπορεί να είναι οτιδήποτε, αρκεί να είναι χρήσιμη έτσι ώστε να μπορέσουμε να ξεχωρίσουμε κάποια δεδομένα από κάποια άλλα. Για παράδειγμα σε ένα email μια ετικέτα μπορεί να χαρακτηρίζει αν το μήνυμα είναι ανεπιθύμητο (*spam*). Η επιβλεπόμενη μάθηση είναι δημοφιλής για την επίλυση προβλημάτων κατηγοριοποίησης (*classification*) και παλινδρόμησης (*regression*) [73].

Ένα από τα προβλήματα της κατηγοριοποίησης θεωρείται το πρόβλημα στο οποίο οι τιμές εξόδου είναι κατηγορίες και οι τιμές τους είναι διακριτές [53]. Για παράδειγμα αν θέλαμε να ξεχωρίσουμε κοσμήματα βάσει του μετάλλου με τα οποία είναι κατασκευασμένα (έστω ότι έχουμε μόνο χρυσό και ασήμι) τότε θα φτιάχναμε μια ομάδα (κλάση) στην οποία θα ανήκαν τα κοσμήματα που είναι φτιαγμένα από χρυσό και μια άλλη ομάδα στην οποία θα βρίσκονται τα κοσμήματα που είναι από ασήμι.

Στα προβλήματα παλινδρόμησης από την άλλη, οι τιμές εξόδου είναι πραγματικές συνεχής τιμές. Επίσης σε αντίθεση με την κατηγοριοποίηση, εδώ προβλέπουμε την τιμή εξόδου ενώ στην κατηγοριοποίηση απλά ομαδοποιούμε τις εξόδους. Στις μεθόδους παλινδρόμησης προσπαθούμε να βρούμε την συνάρτηση που αντιστοιχίζει μια τιμή εισόδου σε μια τιμή εξόδου και να προβλέψουμε τις τιμές εξόδου για κάποιες συγκεκριμένες τιμές εισόδου [72].

Οι πιο δημοφιλείς μέθοδοι που χρησιμοποιούνται στην επιβλεπόμενη μάθηση κατά την βιβλιογραφία [45][72][74][75][76] είναι:

- Γραμμική παλινδρόμηση (*Linear regression*) [77].
- Δένδρα απόφασης (*Decision trees*) [78][79].
- Μηχανές Διανυσμάτων Υποστήριξης (*Support vector machine-SVM*) [80][81].
- Κ-Κοντινότερων Γειτόνων (*k-Nearest Neighbors*) [82][83][84].
- Ταξινομητές Naïve Bayes (*Naive Bayes classifiers*) [85][86].
- Τυχαία Δάση (*Random forest*) [87][88][89].
- Νευρωνικά δίκτυα (*Neural Networks*) [90][91][92].
- Πολυεπίπεδο Περσέπτρον (*Multilayer Perceptron-MLP*) [93][94].

### 4.1.2 Μη Επιβλεπόμενη Μάθηση (*Unsupervised Learning*)

Η μη επιβλεπόμενη μάθηση είναι το αντίθετο από την επιβλεπόμενη. Στην μη επιβλεπόμενη μάθηση έχουμε μόνο δεδομένα εισόδου και καθόλου δεδομένα εξόδου, δηλαδή δεν έχουμε απαντήσεις που να αντιστοιχούν σε κάποιες τιμές εισόδου. Σκοπός αυτής της μεθόδου είναι να ανακαλύψει μοτίβα δεδομένων και τα δεδομένα που έχουν ομοιότητες μεταξύ τους να τα κατηγοριοποιήσει σε ομάδες [53]. Ο αλγόριθμος που θα επεξεργαστεί τα δεδομένα το μόνο που κάνει είναι να εξετάσει τα δεδομένα εισόδου, να βρει τις ομοιότητες μεταξύ τους και ανάλογα με τις μεταξύ τους ομοιότητες να τα κατατάξει σε ομάδες. Με αυτήν την μέθοδο μπορούμε να κατηγοριοποιήσουμε δεδομένα που δεν έχουν ετικέτες (*unlabeled*). Η πιο διαδεδομένη τεχνική μη επιβλεπόμενης μάθησης είναι η συσταδοποίησης (*clustering*). Άλλες γνωστές τεχνικές κατά την βιβλιογραφία [72][74][95][96][97] είναι η ανίχνευση ανωμαλιών (*anomaly detection*) ,τα νευρωνικά δίκτυα (*neural networks*), οι κανόνες συσχετίσεων (*association rules*), η μείωση διάστασης (*dimensionality reduction*) και οι χάρτες αυτό-οργάνωσης (*self-organizing maps*).

Οι πιο δημοφιλείς μέθοδοι που χρησιμοποιούνται στην μη επιβλεπόμενη μάθηση είναι:

- Αλγόριθμοι συσταδοποίησης (*clustering*) [98][99]:
    - Κ-μέσων (*K-means*) [100].
    - Χ-μέσων (*X-means*) ο οποίος είναι παραλλαγή του Κ-μέσων με αποτέλεσμα να είναι πιο γρήγορος και πιο αποτελεσματικός για σύνολα δεδομένων μεγάλων διαστάσεων [101].
    - Ιεραρχικής συσταδοποίησης *(hierarchical clustering*) [102].
    - CLIQUE ο οποίος είναι ένας από τους πιο γνωστούς αλγόριθμους της κατηγορίας συσταδοποίησης υποχώρου (subspace clustering) και είναι ιδανικός για σύνολα δεδομένων μεγάλων διαστάσεων [33][103].
    - Μοντέλα ανάμιξης (*mixture models*) [104] [105].
- Ανίχνευσης ανωμαλιών (*anomaly detection*): Όπως αναφέρεται και στο [106] η διαδικασία της ανίχνευσης ανωμαλιών χρησιμοποιεί διαφορετικές τεχνικές και αλγορίθμους ανάλογα με το που εφαρμόζεται. Για παράδειγμα για ανίχνευση ανωμαλιών στο τομέα της υγείας (ιατρικά αρχεία κ.λ.π) οι τεχνικές που χρησιμοποιούνται είναι τα νευρωνικά δίκτυα, τα δίκτυα *Bayes*, οι τεχνικές κοντινότερων γειτόνων, τα παραμετρικά στατιστικά μοντέλα και οι τεχνικές που βασίζονται στους κανόνες συστημάτων (*rule based*). Όπως βλέπουμε οι τεχνικές συσταδοποίησης δεν είναι δημοφιλής σε αυτόν τον κλάδο. Από την άλλη για ανίχνευση ανωμαλιών σε πληροφορικά σύστημα (π.χ. *Intrusion Detection Systems*) και σε δίκτυα αισθητήρων για παράδειγμα οι τεχνικές και αλγόριθμοι συσταδοποίησης είναι αυτοί που κυριαρχούν.

- Νευρωνικά δίκτυα (n*eural networks*): Στα νευρωνικά δίκτυα οι πιο διαδεδομένοι αλγόριθμοι είναι οι παρακάτω:
    - Αλγόριθμος μάθησης Hebbian [107].
    - DCGANs (*Generative Adversarial Networks*) [108].
- Κανόνες συσχετίσεων (*association rules*): Ο δημοφιλέστερος αλγόριθμος αυτής της προσέγγισης όπως αναφέρεται στο [95] είναι ο αλγόριθμος Apriori [41].
- Μείωση διάστασης (*dimensionality reduction*): Χρήση τέτοιων μεθόδων έχει ως σκοπό την μείωση κάποιων παραμέτρων στο μοντέλο που εφαρμόζονται έτσι ώστε να μειωθεί το υπολογιστικό κόστος [109]. Οι πιο γνωστοί αλγόριθμοι αυτής της μεθόδου είναι [110]:
    - Ανάλυση σε κύριες συνιστώσες (*Principal component analysis-PCA*).
    - Γραμμική ανάλυση σε κύριες συνιστώσες (*Linear discriminant analysis-LDA*).
- Χάρτες αυτό-οργάνωσης (*self-organizing maps-SOM*): Οι χάρτες αυτό-οργάνωσης είναι χρήσιμοι έτσι ώστε να οπτικοποιήσουμε δεδομένα τα οποία βρίσκονται σε ένα πολυδιάστατο χώρο με την χρήση ενός χώρου με λιγότερες διαστάσεις [72][111][112]. Ουσιαστικά προσπαθούμε να απλοποιήσουμε τον πολυδιάστατο χώρο και να προβάλουμε όλες τις διαστάσεις του σε ένα χώρο με λιγότερες διαστάσεις.

### 4.1.3 Ημί-επιβλεπόμενη μάθηση (*semi supervised learning*)

Η ημί-επιβλεπόμενη μάθηση είναι μια κατηγορία μάθησης η οποία μπορούμε να πούμε ότι βρίσκεται μεταξύ της επιβλεπόμενης και της μη επιβλεπόμενης. Ο λόγος για τον οποίο ονομάστηκε έτσι είναι διότι σε αυτήν την κατηγορία γίνεται χρήση δεδομένων με ετικέτες (*labeled data*) αλλά και δεδομένων χωρίς ετικέτες (*unlabeled data*). Από το σύνολο τον δεδομένων που χρησιμοποιούνται η πλειοψηφία των δεδομενών είναι χωρίς ετικέτες. Η χρήση μερικών δεδομένων με ετικέτες γίνεται μόνο και μόνο για να μπορέσει να δημιουργηθεί μια αποδοτικότερη συνάρτηση ή ένας πληρέστερος ταξινομητής (*classifier*) [113].

Οι μέθοδοι που χρησιμοποιούνται ευρέως και συχνότερα σε αυτήν την κατηγορία είναι [114]:

- EM αλγόριθμος με μοντέλα γενετικής μίξης (*expectation–maximization (EM) algorithm with generative mixture models*) [115].
- Μέθοδοι βασιζόμενοι σε γραφική παράσταση (*Graph-based methods*) [116].
- Αυτό-εκπαιδευόμενη μέθοδος (*self-training method*) [117][118][119].
- Συν-εκπαιδευόμενη μέθοδος (*co-training method*) [120][121].
- Μεταβιβαστικές μηχανές διανυσματικής υποστήριξης (*transductive support vector machines*) [122][123].

### 4.1.4 Ενισχυτική μάθηση (*Reinforcement learning*)

Στην ενισχυτική μάθηση ο αλγόριθμος μαθαίνει από τις ενέργειες που πραγματοποιούνται σε ένα περιβάλλον και την αλληλεπίδραση που έχει με αυτό. Ο αλγόριθμος λαμβάνει ανάδραση από το περιβάλλον για κάθε απόφαση του. Η ανάδραση αυτή έχει την μορφή ανταμοιβής αν κάνει σωστή επιλογή και ποινής αν κάνει λάθος επιλογή, χωρίς όμως να του αποκαλύπτεται η σωστή απάντηση για κάθε του επιλογή. Μετά από μια σειρά δοκίμων ο αλγόριθμος θα πρέπει να μάθει την ακολουθία σωστών επιλογών έτσι ώστε να βρει το σωστό αποτέλεσμα [124]. Αυτήν η μέθοδος μάθησης έχει αρκετές εφαρμογές κυρίως στην μάθηση επιτραπέζιων παιχνιδιών και στα ρομπότ. Για παράδειγμα σε ένα παιχνίδι σκάκι ο παίχτης κινείται στην σκακιέρα κάνοντάς κινήσεις που πιστεύει ότι θα του φέρουν την

νίκη και αναπροσαρμόζει τις κινήσεις του ανάλογα με την στρατηγική που ακολουθεί ο αντίπαλός του [71]. Άλλο ένα παράδειγμα είναι το παράδειγμα του ρομπότ στον λαβύρινθο. Σε αυτό το παράδειγμα το ρομπότ προσπαθεί να βρει την έξοδο κάνοντας επανειλημμένες προσπάθειες εκ των οποίων κάποιες είναι επιτυχημένες και κάποιες αποτυχημένες. Μετά από αρκετές προσπάθειες βρίσκει την έξοδο και μαθαίνει την ακολουθία του μονοπατιού για την έξοδο ύστερα από μια σειρά δοκιμών [53][125].

Στην ενισχυτική μάθηση υπάρχουν δυο δημοφιλής προσεγγίσεις και είναι οι εξής:

- Απευθείας εύρεση πολιτικής (*Direct policy search*): Για την εύρεση πολιτικών χρησιμοποιούνται μέθοδοι που βασίζονται σε βαθμίδες (*gradient-based*) και δημοφιλής αλγόριθμοι αυτής της προσέγγισης είναι οι REINFORCE και ο TD(λ) [126].
- Μάθηση βασιζόμενη στην συνάρτηση τιμών (*Value function based learning*): Αυτές οι προσεγγίσεις βασίζονται στην θεωρία των Μαρκοβιανών διαδικασιών απόφασης (*Markov decision processes_MPDs*) για εκτίμηση της πολιτικής που θα οδηγήση σε καλύτερα αποτελέσματα [124]. Αυτήν η προσέγγιση περιλαμβάνει δύο είδη μεθόδων οι οποίες είναι [53][127]:
    - Μέθοδοι Μόντε Κάρλο
    - Μέθοδοι χρονικής διαφοράς (*Temporal Difference methods*). Οι πιο δημοφιλής αλγόριθμοι αυτής της κατηγορίας είναι οι TD(0), Sarsa και ο Q-learning.


*Εικόνα 1. Ταξινόμηση αλγορίθμων μηχανικής μάθησης [131].*

### 4.1.4 Εξελικτική μάθηση (*Evolutionary Learning*)

Η εξελικτική μάθηση βασίζει την θεωρία της στην βιολογία και στο πως οι οργανισμοί αναπαράγονται και εξελίσσονται καθώς μεγαλώνουν ηλικιακά. Οι μηχανισμοί που έχουν αναπτυχθεί στην εξελικτική μάθηση είναι εμπνευσμένοι από διαδικασίες την βιολογικής εξέλιξης όπως η αναπαραγωγή, η

μετάλλαξη, η αναπροσαρμογή και η επιλογή [128]. Στην εξελικτική μάθηση κυριαρχούν οι γενετικοί αλγόριθμοι (*Genetic Algorithms-GA*). Αυτοί αποτελούν μία από τις μεθόδους που χρησιμοποιούνται στην μηχανική μάθηση και υλοποιούν την λογική της εξελικτικής μάθησης. Οι γενετικοί αλγόριθμοι λειτουργούν όπως λειτουργεί η αναπαραγωγική διαδικασία των βιολογικών οργανισμών καθώς αυτοί εξελίσσονται και προσαρμόζονται στο περιβάλλον. Από την επιστήμη της ιατρικής και της βιολογίας είναι γνωστό ότι αυτήν η διαδικασία λαμβάνει χώρα στα χρωμοσώματα τα οποία καταγράφουν την δομή κάθε οργανισμού από τα γονίδια του [129].

Για να μπορέσουμε να δημιουργήσουμε γενετικούς αλγορίθμους και να τους μοντελοποιήσουμε πρέπει συμφώνα με το [130] να ακολουθήσουμε τα εξής στάδια:

1. Δημιουργία του αρχικού πληθυσμού τυχαία. Έτσι θα αναπαραστήσουμε τα προβλήματα που αντιμετωπίζουμε σαν χρωμοσώματα.
2. Πρέπει να βρούμε ένα τρόπο έτσι ώστε να υπολογίσουμε την προσαρμοστικότητα κάθε λύσης που βρίσκουμε στο περιβάλλον. Δηλαδή θα πρέπει να βρούμε ένα τόπο για να αξιολογούμε πόσο καλά προσαρμόζεται στο περιβάλλον, κάθε στοιχείο του πληθυσμού που δημιουργήσαμε στο προηγούμενο στάδιο.
3. Ακολουθούμε τα βήματα αναπαραγωγής τα οποία είναι:
    I. Επιλογή των καλυτέρων στοιχείων (αυτά με την καλύτερη προσαρμοστικότητα) του συνόλου για αναπαραγωγή. Αυτά θα είναι οι γονείς.
    II. Δημιουργία ενός ζευγαριού γονέων έτσι ώστε να παράγουν απογόνους.
    III. Αξιολόγηση της προσαρμοστικότητας στο περιβάλλον των νέων οντοτήτων που γεννήθηκαν.
4. Αντικατάσταση των οντοτήτων του παλιού πληθυσμού που δεν ταιριάζουν πλήρως στο περιβάλλον με νέες οντότητες που ταιριάζουν καλύτερα.
5. Μετάβαση στο στάδιο 2 και επανάληψη των βημάτων.

Πρέπει να επαναλάβουμε αυτά στάδια από το στάδιο 2 και μετά, όσες φορές θέλουμε, έτσι ώστε να βρούμε τον καταλληλότερο πληθυσμό.

## 4.2 Βασικά βήματα της διαδικασίας της μηχανικής μάθησης

Οι μέθοδοι και οι αλγόριθμοι της μηχανικής δεν μπορούν να εφαρμοστούν απευθείας σε μια οποιαδήποτε συλλογή δεδομένων. Για να μπορέσουμε να εξάγουμε σωστά αποτελέσματα και πληροφορίες από τα δεδομένα, πρέπει να τηρηθεί μια διαδικασία, έτσι ώστε να μπορέσουμε να εξάγουμε τα καλύτερα δυνατά αποτελέσματα και η πληροφορία που θα πάρουμε να είναι χρήσιμη. Τα βήματα αυτής της διαδικασίας περιγράφονται συνοπτικά παρακάτω.

1. <u>Ορισμός και κατανόηση του προβλήματος</u>: Αυτό είναι το βασικότερο βήμα από όλα τα υπόλοιπα. Το πρόβλημα θα πρέπει να οριστεί επακριβώς και να γίνει πλήρης κατανόηση του, μόνο έτσι θα μπορέσουμε να καταλάβουμε το τι αντιμετωπίζουμε και τι αποτελέσματα θα πρέπει να περιμένουμε.
2. <u>Ανάλυση των δεδομένων</u>: Πρέπει να γίνει επιλογή ενός υποσυνόλου δεδομένων από όλα αυτά που έχουμε στην διάθεση μας και να ξεκαθαριστεί τι μπορούμε να αναλύσουμε αποτελεσματικά και τι όχι. Το να χρησιμοποιήσουμε όλα τα δεδομένα που έχουμε στην διάθεση μας δεν είναι καλή πρακτική. Αυτό διότι πολλά από αυτά τα δεδομένα που έχουμε πιθανόν να διαφέρουν μεταξύ τους και να μην μπορούν να αναλυθούν με την ίδια μέθοδο μηχανικής μάθησης. Επιπλέον στους αλγόριθμους μηχανικής μάθησης στην πλειοψηφία τους το υπολογιστικό κόστος

αυξάνεται κατά πολύ όταν διαχειρίζονται μεγάλους όγκους δεδομένων. Οπότε καλό είναι να συλλέγουμε όσο λιγότερα δεδομένα είναι δυνατό για την δημιουργία ενός αντιπροσωπευτικού δείγματος του ευρύτερου συνόλου.

3. Προ-επεξεργασία δεδομένων: Αυτό το βήμα είναι συμπληρωματικό του προηγούμενου. Εδώ έχουμε επιλέξει ήδη ένα υποσύνολο δεδομένων το οποίο όμως είναι αρκετά μεγάλο σε όγκο και πιθανόν να περιέχει και αρκετή περιττή πληροφορία. Οπότε προσπαθούμε να αφαιρέσουμε, να τροποποιήσουμε ή να μορφοποιήσουμε για παράδειγμα τυχόν θόρυβο, ακραίες τιμές (outliers), κατεστραμμένα αρχεία , διπλό-εγγραφές και γενικά ότι θεωρείται περιττό και μπορεί να έχει επηρεάσει αρνητικά την όλη διαδικασία.
4. Επιλογή χαρακτηριστικών (*feature selection*): Η επιλογή χαρακτηριστικών τα τελευταία χρόνια αποτελεί αναπόσπαστο κομμάτι της εξόρυξης δεδομένων και τη μηχανικής μάθησης. Γενικά όπου έχουμε μεγάλο όγκο δεδομένων προς επεξεργασία η επιλογή χαρακτηριστικών βοηθά στην επιλογή των χρησιμότερων χαρακτηριστικών τα οποία θα βοηθήσουν στην επίλυση του προβλήματος που αντιμετωπίζουμε. Με την διαδικασία επιλογής χαρακτηριστικών αφαιρούνται τα διπλότυπα και μη συναφή δεδομένα. Αυτό συνεισφέρει στην δημιουργία και επιλογή ενός αντιπροσωπευτικού υποσυνόλου δεδομένων στο οποίο περιλαμβάνονται μόνο χρήσιμα δεδομένα, έχει μικρότερο όγκο και μεγαλύτερη αξία όσων αφορά την πληροφορία που αυτό έχει.
5. Επιλογή της μεθόδου: Από τα προηγούμενα βήματα ο όγκος των δεδομένων είναι μικρότερος και έχουν απαλειφθεί τα περιττά και άχρηστα δεδομένα. Σε αυτό το βήμα πλέον μπορούμε να εξετάσουμε και να αναλύσουμε τα δεδομένα και να επιλέξουμε ποια μέθοδος μηχανικής μάθησης (π.χ. συσταδοποίηση, νευρωνικά δίκτυα κλπ.) είναι καλύτερη για την επίλυση του προβλήματος που αντιμετωπίζουμε.
6. Επιλογή του αλγορίθμου: Αυτό το βήμα είναι συνέχεια του προηγουμένου και είναι πολύ κρίσιμο διότι πάνω σε αυτόν τον αλγόριθμο βασίζεται η λειτουργία του συστήματος. Η κατανόηση των μαθηματικών αρχών και του τρόπου λειτουργίας του αλγορίθμου είναι πολύ σημαντική και πρέπει να γίνει ύστερα από αρκετή μελέτη.
7. Εκπαίδευση του συστήματος: Έχοντας το σύνολο δεδομένων , την μέθοδο και τον αλγόριθμο τώρα πρέπει να εκπαιδεύσουμε το σύστημα που υλοποιήθηκε για να μάθει από τα δεδομένα εισόδου που του εισάγουμε, έτσι να μπορεί να προβλέπει τιμές εξόδου και να εξάγει αποτελέσματα.
8. Αξιολόγηση του συστήματος: Η αξιολόγηση του συστήματος περιλαμβάνει μια σειρά πειραμάτων με διαφορετικά σύνολα δεδομένων σαν είσοδο. Από αυτήν την σειρά πειραμάτων θα πρέπει τα αποτελέσματα, δηλαδή οι τιμές εξόδου που παράγει, να ελεγχθούν εξονυχιστικά και να γίνει η αξιολόγησή τους. Με άλλα λόγια σε αυτό το τελευταίο βήμα γίνεται η αξιολόγηση της ακρίβειας και κατά συνέπεια της αποδοτικότητας του συστήματος. Όταν η ποιότητα και η ακρίβεια των αποτελεσμάτων της διαδικασίας της αξιολόγησης δεν είναι ικανοποιητικά, οι σχεδιαστές του συστήματος, πραγματοποιούν αλλαγές και τροποποιήσεις ενός ή περισσότερων βημάτων της όλης διαδικασίας, μέχρι να επιτύχουν τα επιθυμητά αποτελέσματα.


*Εικόνα 2. Βασικά βήματα της διαδικασίας της μηχανικής μάθησης*

## 4.3 Σύνοψη

Σε αυτήν την ενότητα έγινε μια σύντομη αναφορά των μεθόδων και αλγορίθμων που χρησιμοποιούνται για μηχανική μάθηση. Σκοπός ήταν να δοθεί μια σφαιρική γνώση γύρω από την μηχανική μάθηση χωρίς να εισέλθουμε εις βάθος σε κάθε μια από τις μεθόδους και τους αλγορίθμους που αναλύσαμε διότι αυτό ξεφεύγει από τον σκοπό της παρούσας διπλωματικής. Τέλος αναφερθήκαμε στα βασικά βήματα που πρέπει να τηρηθούν κατά την ανάπτυξη ενός τέτοιου συστήματος.

Η επιστημονική κοινότητα ασχολείται χρόνια με το πώς θα κάνει ένα μηχάνημα να μαθαίνει από μόνο του και να παίρνει μόνο του αποφάσεις. Από τις πρώτες προσπάθειες και θεωρίες, μέχρι τις μέρες μας, έχει γίνει θεαματική πρόοδος και η υλοποίηση συστημάτων με στοιχεία τεχνικής νοημοσύνης είναι πλέον πραγματικότητα. Το μέλλον ανήκει σε τέτοια συστήματα τα οποία θα λειτουργούν σχεδόν αυτόνομα και η ανθρώπινη παρέμβαση θα είναι απαραίτητη μόνο για τις κρίσιμες αποφάσεις, τις οποίες μόνο η ανθρώπινη νοημοσύνη μπορεί να διαχειριστεί.

# Κεφάλαιο 5ο

## Ανίχνευση Ανωμαλιών

Η έρευνα για την ανάπτυξη αποδοτικότερων συστημάτων ανίχνευσης εισβολών έχει πάρει μεγάλες διαστάσεις τα τελευταία χρόνια λόγω του ότι οι επιθέσεις προς τα πληροφορικά συστήματα αυξάνονται συνεχώς και τα ποσοστά επιτυχίας τους είναι αρκετά υψηλά. Επιπλέον οι άνθρωποι που πραγματοποιούν αυτές τις επιθέσεις δεν είναι μόνο άνθρωποι που ασχολούνται ερασιτεχνικά με την ανάπτυξη και εξέλιξη νέων μορφών επιθέσεων. Οι περισσότερες από τις επιθέσεις που πραγματοποιήθηκαν σε μεγάλα συστήματα και κρίσιμες υποδομές κρύβουν από πίσω τους μεγάλες ομάδες καταρτισμένων ανθρώπων οι οποίοι έχουν ως κύριο επάγγελμα τους την ανάπτυξη νέων μεθόδων παραβίασης πληροφοριακών συστημάτων [132]. Αυτοί οι νέα γενιά χάκερ αυξάνεται συνεχώς και τους έχει δοθεί η ονομασία κυβερνό-εγκληματίες διότι διαπράττουν εγκλήματα μέσω του διαδικτύου τα οποία τις περισσότερες φορές του αποφέρουν και μεγάλα χρηματικά ποσά. Όπως καταλαβαίνουμε το κύβερνο-έγκλημα στις μέρες είναι ένας χώρος που προσελκύει αρκετούς επιδέξιους επαγγελματίες του χώρου της πληροφορικής και με πολύ υψηλές χρηματικές απολαβές σε σχέση με τα κλασσικά επαγγέλματα σε αυτόν τον χώρο [133].

## 5.1 Είδη εισβολών και εισβολέων

Σύμφωνα με το [134] προσπάθεια εισβολής θεωρείται κάθε προσπάθεια μη εξουσιοδοτημένης πρόσβασης με σκοπό την πρόσβαση, την υποκλοπή ή την αλλοίωση πληροφοριών. Επιπλέον εισβολή θεωρείται κάθε προσπάθεια που έχει σκοπό να καταστήσει ένα σύστημα άχρηστο ή να επηρεάσει την αξιοπιστία του. Για παράδειγμα μια επίθεση τύπου *DDoS* έχει σκοπό να εξαντλήσει τους πόρους ενός συστήματος και να το θέσει εκτός διαθεσιμότητας, ενώ μια επίθεση που εξαπολύει έναν ιό ή ένα κακόβουλο λογισμικό (*malware*) εκμεταλλεύεται μια ευπάθεια του συστήματος με σκοπό να μολύνει ένα σύστημα επηρεάζοντας την ακεραιότητα και εμπιστευτικότητα του. Υπάρχουν πολλών ειδών επιθέσεις

που πρέπει να αντιμετωπιστούν και να χτιστούν άμυνες για αυτές. Τα είδη των επιθέσεων και των εισβολών ποικίλουν [135][136][137], περιληπτικά οι τύποι των επιθέσεων είναι οι εξής:

- Ιός (*virus*): Ένα κακόβουλο λογισμικό το οποίο εγκαθίσταται σε ένα σύστημα χωρίς την γνώση και την άδεια του χρήστη.
- Σκουλήκι (*worm*): Κακόβουλο λογισμικό το οποίο μεταδίδεται κυρίως μέσω του διαδικτύου και έχει την δυνατότητα να μεταπηδά από μηχάνημα σε μηχάνημα με σκοπό να μολύνει όσο περισσότερα μηχανήματα μπορεί.
- Δούρειος Ίππος (*Trojan*): Σκοπός των λογισμικών αυτών είναι η δημιουργία πίσω πόρτας (*backdoor*) έτσι ώστε να επιτρέψει τον επιτιθέμενο να έχει μη εξουσιοδοτημένη πρόσβαση σε ένα σύστημα.
- Άρνηση Υπηρεσίας (*DoS*): Επιθέσεις τέτοιου τύπου έχουν ως στόχο να εξαντλήσουν τους υπολογιστικούς πόρους ενός πληροφοριακού συστήματος ή να καταναλώσουν όλη το διαθέσιμο εύρος (*bandwidth*) ενός δικτύου και να επηρεάσουν την διαθεσιμότητα του στους εξουσιοδοτημένους χρήστες.
- Επιθέσεις δικτύου (*Network Attacks*): Αυτού του είδους οι επιθέσεις στοχεύουν στον να υποβιβάσουν την ασφάλεια του δικτύου με οποιοδήποτε τρόπο. Τέτοιες επιθέσεις εκμεταλλεύονται κυρίως ευπάθειες των πρωτοκόλλων του διαδικτύου.
- Φυσικές επιθέσεις: Είναι η προσπάθεια πρόκλησης καταστροφών στα φυσικά μηχανήματα ή υποδομές.
- Επιθέσεις κωδίκων πρόσβασης (*password attacks*): Στόχος αυτών των επιθέσεων είναι η εύρεση κωδικών πρόσβασης εξουσιοδοτημένων χρηστών.
- Συγκέντρωση πληροφοριών (*information gathering*): Η συγκέντρωση πληροφοριών αποσκοπεί στην συλλογή πληροφοριών για ένα πληροφοριακό σύστημα και τα τμήματα λογισμικού από τα οποία αυτό απαρτίζεται, έτσι ώστε να γίνει έλεγχος για διάφορες ευπάθειες που μπορεί να έχουν και πιθανά τρωτά σημεία.
- U2R (*user to root*): Σε αυτού του είδους τις επιθέσεις εκμεταλλευόμενες ο επιτιθέμενος κάποιες ευπάθειες του συστήματος μπορεί από απλός χρήστης που είναι να αποκτήσει δικαιώματα διαχειριστή.
- R2L (*remote to local*): Τέτοιου είδους επιθέσεις αναφέρονται στην απομακρυσμένη επικοινωνία ενός κακόβουλου χρήστη με ένα πληροφοριακό σύστημα χωρίς να είναι εξουσιοδοτημένος. Ο επιτιθέμενος εδώ μπορεί να διαχειριστεί λειτουργίες του συστήματος έχοντας πρόσβαση απομακρυσμένα εκμεταλλευόμενος κάποιες πίσω πόρτες του συστήματος ή κάποια λογισμικά που έχουν αλλοιωθεί με κακόβουλο κώδικα και έχουν εγκατασταθεί στο σύστημα στόχο.
- Probe: Η διαδικασία αυτή της επίθεσης περιλαμβάνει εκτενή σάρωση του δικτύου του συστήματος που είναι στόχος, για εντοπισμό των μηχανήματων που βρίσκονται σε αυτό και γενικότερα όσο περισσότερων πληροφοριών γίνεται. Αυτήν η διαδικασία γίνεται με ειδικά εργαλεία (*IP scanners , portscanner* κ.λ.π) τα οποία πληροφορούν τον επιτιθέμενο για το είδος των συσκευών που υπάρχουν στο δίκτυο. Από αυτό ο επιτιθέμενος μπορεί να αντλήσει πληροφορίες και να αναζητήσει βάσει των πληροφοριών που έχει στην κατοχή του, ποιες πιθανές ευπάθειες μπορεί να υπάρχουν στο πληροφοριακό σύστημα, έτσι ώστε να τις εκμεταλλευτεί για να εισβάλει.

Τα είδη των εισβολέων σύμφωνα με το [134] είναι δύο οι εξωτερικοί εισβολείς και οι εσωτερικοί. Οι εξωτερικοί εισβολείς είναι αυτοί οι οποίοι δεν έχουν καμία εξουσιοδοτημένη πρόσβαση στο σύστημα

και προσπαθούν να βρούνε τρόπους έτσι ώστε να αποκτήσουν πρόσβαση. Αυτοί τις περισσότερες φορές γνωρίζουν λίγα πράγματα για την δομή του πληροφοριακού συστήματος και μέσω διαδικασιών όπως το η συλλογή πληροφοριών, προσπαθούν να συλλέξουν όσες περισσότερες πληροφορίες μπορούν για το σύστημα και να ανακαλύψουν πιθανές αδυναμίες του.

Οι εσωτερικοί εισβολείς από την άλλη είναι χρήστες του πληροφοριακού συστήματος οι οποίοι έχουν πρόσβαση σε κάποια από τα τμήματα του συστήματος και προσπαθούν να αποκτήσουν περισσότερα δικαιώματα όπως για παράδειγμα δικαιώματα διαχειριστή. Αυτού του είδους οι εισβολείς έχουν μερική ως πολύ καλή γνώση του συστήματος και είναι συνήθως υπάλληλοι του οργανισμού στον οποίο ανήκει το πληροφοριακό σύστημα. Για αυτό τον λόγο είναι και πιο επικίνδυνοι από τους εξωτερικούς εισβολείς διότι πρώτον έχουν άμεση πρόσβαση στο σύστημα και δεύτερον γνωρίζουν τις κρίσιμες υποδομές του πληροφοριακού συστήματος και μπορούν να προκαλέσουν μεγάλες ζημιές. Επίσης έχουν την δυνατότητα να καλύψουν τα ίχνη τους καλυτέρα από ένα εξωτερικό εισβολέα διότι μπορούν να χρησιμοποιήσουν λογαριασμούς άλλων χρηστών για να εισέλθουν παράνομα στον σύστημα και να μην γίνουν αντιληπτοί. Επιπρόσθετα οι εσωτερικοί εισβολείς έχουν και το πλεονέκτημα του χρόνου σε σχέση με τους εξωτερικούς. Ο εσωτερικός εισβολέας ξέρει τα μηχανήματα και την υποδομή του πληροφοριακού συστήματος και δεν χρειάζεται να δαπανήσει χρόνο για να συλλέξει πληροφορίες και να αποκτήσει πρόσβαση στο σύστημα. Από την άλλη ο εξωτερικός πρέπει να σπαταλήσει αρκετό χρόνο για αυτήν την διαδικασία και έπειτα να βρει τρόπο να εισβάλει. Για αυτό τον λόγο οι εσωτερικοί επιτιθέμενοι θεωρούνται αρκετά πιο επικίνδυνοι από τους εξωτερικούς επιτιθέμενους και είναι δυσκολότερο να τους περιορίσουμε διότι έχουν ήδη πρόσβαση σε πόρους του συστήματος αλλά και πληροφορίες για αυτό.

Ένα παράδειγμα εσωτερικού επιτιθέμενου είναι το περιστατικό που συνέβη το 1996 στην εταιρεία ρολογιών *Omega*. Εκεί ένας υπάλληλος την εταιρείας διέγραψε όλα τα αρχεία του *Novell NetWare 3.12 file server* και κατέστρεψε και τα backup αρχεία έτσι ώστε να μη μπορούν να επαναφέρουν τον *server* σε λειτουργία. Όλο αυτό είχε σαν αποτέλεσμα την διακοπή της διαδικασίας παραγωγής για αρκετό χρονικό διάστημα αλλά και την απώλεια σχεδόν όλων των τεχνικών σχεδίων των ρολογιών της εταιρείας που βρισκόταν στον *server*. Υπολογίζεται ότι όλο αυτό κόστισε στην εταιρεία περίπου 10 εκατομμύρια δολάρια χωρίς να μπορούν να υπολογιστούν με ακρίβεια οι ζημιές. Σύμφωνα με τους ειδικούς που εξέτασαν την υπόθεση το πραγματικό κόστος ήταν αρκετά υψηλότερο [138].

## 5.2 Ανίχνευση Ανωμαλιών

Από τις αρχές της δεκαετίας του 90’ αναπτύχθηκαν πολλά συστήματα ανίχνευσης εισβολών. Τα περισσότερα από αυτά είχαν ως στόχο την προστασία των μηχανημάτων (*host*) και λίγα από αυτά είχαν δημιουργήθηκαν για την προστασία του δικτύου (*network IDS*). Στα περισσότερα η τεχνική ανίχνευσης που χρησιμοποιούταν βασιζόταν σε υπογραφές (*signature-based*) δηλαδή η τεχνική ανίχνευσης που χρησιμοποιούσαν ήταν αυτό που αποκαλούμε «τεχνική ανίχνευσης κακής χρήσης» (*misuse detection*) [139]. Όπως αναφέρονται στα [11][20][140] δύο είναι τα κυρίαρχα είδη μεθόδων ανίχνευσης εισβολών. Αυτές είναι η ανίχνευση κακής χρήσης και η ανίχνευση ανωμαλιών. Υπάρχει και ένα τρίτος είδος η υβριδική τεχνική (*hybrid*) η οποία είναι συνδυασμός των δυο παραπάνω [141][142].

Όπως αναφέραμε και στο 2$^{ο}$ Κεφάλαιο η μέθοδος ανίχνευσης κακής χρήσης ψάχνει για μοτίβα απειλών που είναι ήδη γνωστά, έχουν αναλυθεί και τους έχει δοθεί μια υπογραφή . Σε αντίθεση η ανίχνευση ανωμαλιών έχει την δυνατότητα να ανιχνεύσει και απειλές οι οποίες δεν έχουν αναλυθεί προηγουμένως, δηλαδή έχει την ικανότητα να ανιχνεύει ασυνήθιστα μοτίβα συμπεριφοράς που

λαμβάνουν χώρα σε ένα πληροφοριακό σύστημα ή δίκτυο. Αν και τα συστήματα ανίχνευσης εισβολών που βασίζονται στην μέθοδο κακής χρήσης εξελίσσονται συνεχώς και έχουν προσθέσει ακόμα ένα επίπεδο προστασίας στα σημερινά πληροφορικά συστήματα, πλέον δεν είναι όσο αποδοτικά θα έπρεπε. Αυτό συμβαίνει γιατί οι επιθέσεις στην σημερινή εποχή είναι πιο εκλεπτυσμένες και οι σχεδιαστές αυτών των επιθέσεων έχουν βρει τρόπους να παρακάμπτουν τα παραδοσιακά συστήματα που βασίζονται σε υπογραφές. Πολλές από τις σύγχρονες επιθέσεις περνούν σχεδόν απαρατήρητες και αυτό είναι που έχει προβληματίσει την επιστημονική κοινότητα. Για αυτόν τον λόγο οι περισσότεροι επιστήμονες τα τελευταία χρόνια έχουν επικεντρωθεί στην ανάπτυξη συστημάτων τα οποία βασίζονται στην τεχνική της ανίχνευσης ανωμαλιών η οποία, δίνει την δυνατότητα ανίχνευσης γνωστών αλλά και άγνωστων επιθέσεων.

Η ανίχνευση ανωμαλιών προσπαθεί να βρει μοτίβα μέσα στα δεδομένα τα οποία δεν συνάδουν με την συνηθισμένη συμπεριφορά ενός συστήματος. Μέσω της εύρεσης ανωμαλιών από ένα σύνολο δεδομένων μπορούμε να εξάγουμε χρήσιμη πληροφορία την οποία πληροφορία μπορούμε να την αναλύσουμε για να εξάγουμε συμπεράσματα. Για παράδειγμα, ο εντοπισμός ενός ανωμάλου μοτίβου διαδικτυακής κίνησης μπορεί να σημαίνει ότι κάποιος υπολογιστής που έχει παραβιαστεί μπορεί να στέλνει πληροφορίες ευαίσθητων δεδομένων σε έναν απομακρυσμένο μηχάνημα. Με λίγα λόγια από την παραπάνω ανωμαλία που εντοπίσαμε έχουμε εντοπίσει δυο πολύ σημαντικά συμβάντα. Πρώτον ότι ένα μηχάνημα έχει παραβιαστεί και παρατηρείται ασυνήθιστη διαδικτυακή κίνηση. Δεύτερον με περαιτέρω ανάλυση αυτής της κίνησης, παρατηρούμε ότι επικοινωνεί με ένα απομακρυσμένο μηχάνημα στέλνοντας συγκεκριμένες πληροφορίες.

Για να μπορέσει όμως το σύστημα ανίχνευσης εισβολών να ανιχνεύσει ανώμαλες συμπεριφορές πρέπει πρώτα να καταγράψει την συμπεριφορά των χρηστών και της διαδικτυακής κίνησης έτσι ώστε να δημιουργήσει προφίλ φυσιολογικής συμπεριφοράς. Οποιαδήποτε συμπεριφορά διαφοροποιείται από τα προφίλ που έχουν δημιουργηθεί τότε χαρακτηρίζεται ως ανώμαλη [4]. Όσον αφορά τις ανωμαλίες του δικτύου το [143] τις κατατάσσει στις παρακάτω δυο κατηγορίες.

- Ανωμαλίες που σχετίζονται με την απόδοση (*performance related*): Παραδείγματα τέτοιων ανωμαλιών είναι οι αποτυχίες εξυπηρετητών φακέλων (*file server*), οι καταιγίδες αναμετάδοσης (*broadcast storms*) , η δημιουργία συμφόρησης στο δικτύου (*transcient congestion*) και οι φλύαροι κόμβοι (*babbling nodes*) [144][145].
- Ανωμαλίες που σχετίζονται με την ασφάλεια (*security related*): Σε αυτήν την κατηγορία είναι επιθέσεις τύπου *DoS* [146][147] και γενικότερα εισβολές στο δικτύο (*network intrusions*) όπως επιθέσεις πλημμύρας (*flooding attacks*)[148][149].

## 5.3 Ένα γενικό μοντέλο συστήματος ανίχνευσης εισβολών

Κατά τον Axelsson στο [150] μια γενική αρχιτεκτονική ενός συστήματος ανίχνευσης εισβολών πρέπει να περιέχει τα ακόλουθα μέρη (*modules*):

- Έλεγχος της συλλογής δεδομένων (*audit data collection*): Αυτό το κομμάτι του συστήματος θα χρησιμοποιηθεί για την συλλογή δεδομένων. Έπειτα τα δεδομένα θα αναλυθούν από τον αλγόριθμο ανίχνευσης για να βρεθούν ύποπτες δραστηριότητές. Η πηγή των δεδομένων μπορεί να είναι αρχεία καταγραφής τύπου *logs* από τα φυσικά μηχανήματα και τις εφαρμογές, κίνηση δικτύου κ.α.
- Αποθήκευση των δεδομένων ελέγχου (*audit data storage*): Ένα από τα σημαντικότερα κομμάτια του συστήματος είναι αυτό διότι εδώ αποθηκεύονται όλα τα δεδομένα που συλλέγονται. Εδώ

αποθηκεύονται δεδομένα που έχουν ήδη αναλυθεί ή που θα αναλυθούν μεταγενέστερα. Το μεγάλο πρόβλημα που αντιμετωπίζεται είναι ότι ο όγκος των δεδομένων είναι υπέρογκος και πρέπει ο αποθηκευτικός χώρος να είναι επαρκής.

- Δεδομένα ρυθμίσεων (*configuration data*): Το πολυτιμότερο κομμάτι του συστήματος είναι αυτό. Σε αυτό αποθηκεύονται όλες οι ρυθμίσεις και η ευαίσθητη πληροφορία για την λειτουργία του συστήματος όπως ο τρόπος λειτουργίας του αλγορίθμου, πως αυτό ανταποκρίνεται σε εισβολές, κάθε πότε συλλεγεί δεδομένα κ.α.
- Δεδομένα αναφοράς (*reference data*): Σε αυτός τμήμα του συστήματος αποθηκεύονται οι πληροφορίες που αφορούν γνωστές εισβολές. Με λίγα λόγια αποθηκεύονται οι υπογραφές των γνωστών απειλών ή τα προφίλ φυσιολογικής συμπεριφοράς.
- Επεξεργασία δεδομένων (*processing data*): Αυτό το τμήμα του συστήματος είναι υπεύθυνο για την επεξεργασία των δεδομένων και για την εξαγωγή αποτελεσμάτων και πληροφοριών για τις εκάστοτε εισβολές που ανίχνευσε.
- Συναγερμός (*alarm*): Αυτό το τμήμα του συστήματος επεξεργάζεται όλη την πληροφορία που βγαίνει σαν έξοδος μετά την επεξεργασία και είναι υπεύθυνο για την έγερση συναγερμού, οποίος θα ειδοποιήσει τον άνθρωπο, ο οποίος είναι υπεύθυνος ασφαλείας για πιθανή εισβολή και από εκεί και πέρα είναι όλα είναι θέμα χειρισμού, αποφάσεων και ενεργειών του υπευθύνου ατόμου.

Επιπλέον μετά την ανάπτυξη ενός συστήματος ανίχνευσης εισβολών βασιζόμενο στην τεχνική ανίχνευσης ανωμαλιών θα πρέπει το σύστημα αρχικά να παραμετροποιηθεί. Έπειτα πρέπει να εκπαιδευτεί έτσι ώστε να δημιουργήσει τα προφίλ φυσιολογικής συμπεριφοράς του πληροφοριακού συστήματος και των χρηστών. Τέλος θα πρέπει γίνει έλεγχος αποτελεσματικότητας, δηλαδή αν ανιχνεύει απειλές και εγείρει συναγερμούς [**Error! Reference source not found.**]. Σε αυτόν τον τελικό έλεγχο γίνονται κάποιες δοκιμές και βάσει κάποιων κριτήριων ποιότητας που έχουν θέσει οι σχεδιαστές του συστήματος γίνεται η αξιολόγηση του. Υπάρχουν αρκετές τεχνικές αξιολόγησης ενός συστήματος αλλά η πιο διαδεδομένη είναι η τεχνική της σταυρωτής επικύρωσης (*cross validation*) και πιο συγκεκριμένα η *10 fold cross validation* [33][152].

## 5.4 Προϋποθέσεις Ανίχνευσης Ανωμαλιών

Η κεντρική προϋπόθεση της ανίχνευσης ανωμαλιών είναι ότι κάθε δραστηριότητα εισβολής είναι ένα υποσύνολο μιας μη ομαλής δραστηριότητας. Λαμβάνοντας υπόψη ότι ο εισβολέας δεν έχει καμία γνώση για την φυσιολογική και νόμιμη δραστηριότητα των χρηστών, η προσπάθεια εισβολής σε ένα συστήματα κατά πάσα πιθανότητα θα ανιχνευθεί σαν μια μη ομαλή δραστηριότητα. Σε αυτήν την περίπτωση το σύνολο των μη ομαλών δραστηριοτήτων θα είναι το ίδιο με το σύνολο των δραστηριοτήτων που υποδηλώνουν εισβολή. Σε αυτήν την περίπτωση σηματοδοτώντας όλες τις μη ομαλές δραστηριότητες ως εισβολές θα οδηγούσε σε μηδενικό αριθμό *false positives* και *false negatives*. Ωστόσο οι δραστηριότητες που μπορούν να χαρακτηριστούν ως εισβολές δεν συμπίπτουν πάντα μη ομαλές δραστηριότητες. Στο [153] οι συγγραφείς αναφέρουν ότι υπάρχουν τέσσερα πιθανά ενδεχόμενα, των οποίων το καθένα έχει μη μηδενική πιθανότητα:

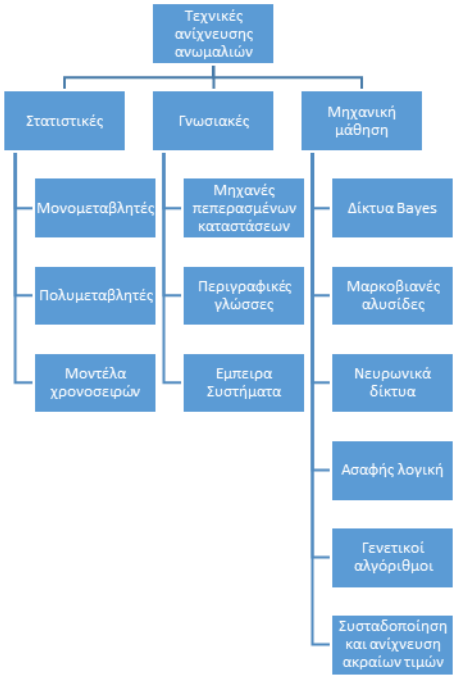
- Εισβολή άλλα όχι ανώμαλη: Αυτό είναι που αποκαλούμε *false negatives.* Σε αυτήν την περίπτωση το σύστημα αποτυγχάνει να ανιχνεύσει μια εισβολή διότι δεν παρουσιάζεται ανώμαλη δραστηριότητα. Έτσι λοιπόν τον σύστημα λανθασμένα δεν ενεργοποιεί συναγερμό για εισβολή.

- Όχι εισβολή αλλά ανώμαλη: Αυτό είναι το *false positives.* Η δραστηριότητα που ανιχνεύεται δεν είναι πράξη εισβολής αλλά επειδή είναι μη ομαλή συμπεριφορά τον σύστημα την ανιχνεύει ως εισβολή. Σε αυτήν την περίπτωση το σύστημα εγείρει συναγερμό ο οποίος είναι αναληθής.
- Όχι εισβολή και όχι ανώμαλη: Χαρακτηρίζονται εν συντομία και ως *true negatives*. Αυτήν η περίπτωση είναι όταν το σύστημα συμπεριφέρεται όπως θα έπρεπε. Δηλαδή δρα ορθά και δεν εγείρει κάποιο συναγερμό επειδή οι δραστηριότητες που παρακολουθεί δεν είναι ούτε ανώμαλες ούτε εισβολές.
- Εισβολή και ανώμαλη: Με άλλα λόγια *true positives*. Εδώ η δραστηριότητα που ανιχνεύει το σύστημα είναι μη ομαλή αλλά και προσπάθεια εισβολής. Σε αυτήν περίσταση το σύστημα εγείρει συναγερμό ορθά.

Σκοπός των σχεδιαστών συστημάτων ανίχνευσης εισβολών είναι να μειώσουν όσο δυνατό γίνεται τον αριθμό των *false negatives* και να περιορίσουν τον αριθμό των *false positives*.

## 5.5 Τεχνικές Ανίχνευσης Ανωμαλιών

Σε αυτήν την ενότητα θα αναφερθούμε περιληπτικά στις βασικές τεχνικές που έχουν μελετηθεί και χρησιμοποιηθεί για την ανίχνευση ανωμαλιών. Οι τεχνικές που έχουν μελετηθεί είναι αρκετές, εμείς εδώ θα αναφερθούμε μόνο στις πιο δημοφιλείς όπως παρουσιάζονται στην σχετική βιβλιογραφία [154][155][156][157][158][159]. Οι τεχνικές αυτές γενικότερα μπορούν να χωριστούν σε τρεις διαφορετικές κατηγορίες βάσει της μεθόδου στην οποία βασίζονται. Έτσι έχουμε τις τεχνικές που βασίζονται στην στατιστική, στην γνώση (*knowledge based*) και στην μηχανική μάθηση.




*Εικόνα 3.* Τεχνικές Ανίχνευσης Ανωμαλιών

### 5.5.1 Στατιστικές μέθοδοι

Η βασική αρχή πάνω στην οποία βασίζονται οι στατιστικές μέθοδοι λέει ότι μια ανωμαλία είναι μια παρατήρηση την οποία υποπτευόμαστε ότι είναι μερικώς ή ολικώς άσχετη, διότι δεν έχει παραχθεί από

το στοχαστικό μοντέλο στο οποίο βασίζουμε τις προβλέψεις μας [160]. Στις στατιστικές μεθόδους που χρησιμοποιούνται για ανίχνευση ανωμαλιών το σύστημα παρατηρεί την δραστηριότητα των υποκειμένων (π.χ. χρήστες, μηχανήματα) και δημιουργεί προφίλ για να αναπαραστήσει την συμπεριφορά τους. Τα προφίλ μου δημιουργεί συνήθως περιλαμβάνουν μετρήσεις, όπως μετρήσεις έντασης, μετρήσεις διανομής αρχείων ελέγχου (*audit records*), κατηγορηματικές μετρήσεις οι οποίες μετρούν την διανομή μια ενέργειας σε διάφορες κατηγορίες και τακτικές μετρήσεις όπως είναι η χρήση της κεντρικής μονάδας επεξεργασίας (*Central Process Unit-CPU*). Τα προφίλ που δημιουργούνται συνήθως είναι δυο. Το ένα είναι το τωρινό προφίλ (*current profile*) και το άλλο είναι αποθηκευμένο προφίλ (*stored profile*). Κατά την επεξεργασία γεγονότων (π.χ. εισερχόμενα πακέτα, αρχεία καταγραφής-*log files*) που συμβαίνουν σε ένα σύστημα ή δίκτυο, το σύστημα ανίχνευσης εισβολών αναβαθμίζει το προσωρινό προφίλ και υπολογίζει περιοδικά μια βαθμολογία μη-ομαλότητας. Αυτήν η βαθμολογία υπολογίζεται συγκρίνοντας το προσωρινό με το αποθηκευμένο προφίλ χρησιμοποιώντας μια συνάρτηση υπολογισμού αντικανονικότητας (*abnormality function*), κάνοντας χρήση όλων των μετρήσεων που έχει συλλέξει για αυτά τα προφίλ. Τέλος αν ο βαθμός μη ομαλότητας είναι υψηλότερος από ένα συγκεκριμένο κατώφλι (*threshold*) τότε το σύστημα εγείρει συναγερμό. Υπάρχουν διάφορες στατιστικές προσεγγίσεις μερικές από αυτές βασίζονται σε μονομεταβλητά (*univariate*)[163], πολυμεταβλητά (*multivariate*) μοντέλα [164][165][166] και μοντέλα χρονοσειρών (*time series*)[167][168][169][170].

Όπως κάθε μέθοδος έτσι και αυτήν έχει κάποια πλεονεκτήματα και κάποια μειονεκτήματα. Τα πλεονεκτήματα των στατιστικών μεθόδων είναι τα εξής:

- Αν οι αρχικές υποθέσεις που κάναμε σχετικά με την κατανομή των δεδομένων είναι αληθής, τότε οι στατιστικές μέθοδοι μας παρέχουν μια στατιστικά τεκμηριωμένη λύση για την ανίχνευση ανωμαλιών.
- Ο βαθμός μη ομαλότητας που δίνεται από μια στατιστική μέθοδο σχετίζεται με το διάστημα εμπιστοσύνης (*confidence interval*), το οποίο μπορεί να χρησιμοποιηθεί ως επιπρόσθετη πληροφορία κατά την διαδικασίας λήψης αποφάσεων σχετικά με το ποιο δείγμα για δοκιμή θα επιλέξουμε.
- Αν η διαδικασία εκτίμησης της κατανομής είναι ανθεκτική σε ανωμαλίες που βρίσκονται στα δεδομένα, τότε οι στατιστικές μέθοδοι μπορούν να χρησιμοποιηθούν σε μη επιβλεπόμενες τεχνικές (*unsupervised*) χωρίς να υπάρχει ανάγκη για δεδομένα με ετικέτες (*labeled data*).
- Οι τεχνικές αυτές δεν χρειάζεται να έχουν προηγουμένη γνώση ευπαθειών ή επιθέσεων για να ανιχνεύσουν εισβολές και ανώμαλες καταστάσεις. Αυτό έχει ως αποτέλεσμα να μπορούν να ανιχνεύσουν τις νέες και τύπου *zero day* εισβολές.
- Παρέχουν ακριβείς ειδοποιήσεις κακόβουλων δραστηριοτήτων διότι αναλύουν αρκετό χρονικό διάστημα την λειτουργία του συστήματος για να φτιάξουν το προφίλ φυσιολογικής συμπεριφοράς. Έτσι αν κάτι μη αναμενόμενο συμβεί το σύστημα θα εγερθεί αμέσως συναγερμός. Αυτό είναι πολύ θετικό για την ανίχνευση επιθέσεων *DoS*. Όταν συμβαίνει μια τέτοιου είδους επίθεση συνήθως η διαδικτυακή κίνηση του αυξάνεται απότομα σε πολύ σύντομο χρονικό διάστημα. Το στατιστικό μοντέλο έχοντας τα προφίλ συμπεριφοράς προηγουμένων χρονικών περιόδων και συγκρίνοντάς το με το τωρινό προφίλ (στο οποίο συμβαίνει η επίθεση) θα αντιληφθεί αμέσως ότι κάτι μη ομαλό συμβαίνει.

Από την άλλη υπάρχουν και αρκετά μειονεκτήματα τα οποία είναι:

- Ένα σημαντικό μειονέκτημα αυτών των μεθόδων είναι ότι βασίζονται στην υπόθεση ότι τα δεδομένα παράγονται από μια συγκεκριμένη κατανομή. Συνήθως αυτήν υπόθεση δεν αληθεύει ιδίως σε δεδομένα μεγάλων διαστάσεων (*high dimensional*) και πραγματικά (δηλαδή όχι τεχνητά παραγόμενα σύνολα δεδομένων) σύνολα δεδομένων της.
- Ακόμα και όταν η υπόθεση είναι λογικά τεκμηριωμένη, υπάρχουν αρκετές υποθέσεις δοκιμής στατιστικών οι οποίες μπορούν να εφαρμοστούν για την ανίχνευση ανωμαλιών. Η επιλογή της καλύτερης στατιστικής μεθόδου συνήθως είναι μια περίπλοκη διαδικασία. Συγκεκριμένα η δημιουργία δοκιμαστικών υποθέσεων για συνθέτες κατανομές οι οποίες πρέπει να ταιριάζουν σε σύνολα δεδομένων μεγάλων διαστάσεων είναι μια τετριμμένη και δύσκολη διαδικασία [161].
- Ο προσδιορισμός κατωφλίων (*thresholds*) για να θέσουμε τα όρια μεταξύ ομαλής και μη ομαλής συμπεριφοράς είναι μια περίπλοκη διαδικασία και χρειάζεται προσοχή κατά την προσπάθεια ισορρόπησης του συστήματος μεταξύ *false negatives* και *false positives*.
- Δεν μπορούμε να μοντελοποιήσουμε όλες τις συμπεριφορές βάσει στατιστικών μεθόδων.
- Η πλειοψηφία των στατιστικών μεθόδων απαιτεί την αρχική υπόθεση μιας ψευδοστατικής (*quasi stationary*) διεργασίας [162]. Όμως αυτήν την υπόθεση δεν μπορούμε να την κάνουμε για τα περισσότερα από τα δεδομένα τα οποία επεξεργάζονται από τα συστήματα ανίχνευσης ανωμαλιών.

### 5.5.2 Μέθοδοι που βασίζονται στην γνώση

Αυτοί οι μέθοδοι συχνά αναφέρονται και ως μέθοδοι εξόρυξης δεδομένων. Σε αυτές τις μεθόδους τα γεγονότα (*events*) που συμβαίνουν σε ένα δίκτυο ή ένα φυσικό μηχάνημα (*host*) ελέγχονται για να δούμε αν ταιριάζουν με συγκεκριμένους προκαθορισμένους κανόνες ή μοτίβα επιθέσεων. Σκοπός είναι η αναπαράσταση μιας γνωστής επίθεσης σε μια πιο γενικευμένη μορφή, για χρήση αυτής της γνώσης σε μελλοντικές επιθέσεις παρόμοιου τύπου. Εδώ η εξόρυξη δεδομένων μπορεί να βοηθήσει να βελτιωθεί η διαδικασία της ανίχνευσης εισβολών έτσι ώστε να μπορέσουμε να επικεντρωθούμε στην ανίχνευση ανωμαλιών. Αναγνωρίζοντας τα όρια μεταξύ ομαλής και μη ομαλής δραστηριότητάς η εξόρυξη δεδομένων βοηθά τον ειδικό αναλυτή να ξεχωρίζει τις ύποπτες δραστηριότητες από την καθημερινή δραστηριότητα του δικτύου. Δημοφιλείς μέθοδοι που χρησιμοποιούνται είναι οι μηχανές πεπερασμένων καταστάσεων (*Finite State Machines-FSMs*)[171][172], οι περιγραφικές γλώσσες (*Description Languages*) [157] και τα έμπειρα συστήματα (*Expert Systems*)[173][174][175][176]. Τα βασικά πλεονεκτήματα αυτών των μεθόδων είναι τα κάτωθι:

- Είναι ανθεκτικές, ευέλικτες και με δυνατότητα κλιμάκωσης τεχνικές.
- Παρουσιάζουν υψηλό ποσοστό ανιχνευσιμότητας αν έχει αναλυθεί σωστά η ομαλή κίνηση και οι απειλές οι απειλές που έχουν κατά καιρού εμφανιστεί.

Από την άλλη κάποια βασικά μειονεκτήματα είναι τα εξής:

- Η ανάπτυξη ποιοτικής γνώσης είναι δύσκολη και χρονοβόρα διαδικασία.
- Αυτές οι μέθοδοι είναι ευάλωτοι σε λανθάνοντες συναγερμούς (*false alarms*) διότι δεν υπάρχει μια διακριτή γραμμή μεταξύ ομαλής και μη ομαλής συμπεριφοράς.
- Μια τέτοια μέθοδος ίσως δεν μπορεί να ανακαλύψει επιθέσεις που συμβαίνουν σπάνια ή είναι εντελώς άγνωστες.
- Η ανανέωση των κανόνων και της γνωσιακής βάσεις (όπου αποθηκεύονται οι αναλύσεις που γίνονται κατά καιρούς) είναι μια διαδικασία συνηθώς με υψηλό κόστος.

### 5.5.3 Μέθοδοι που βασίζονται στην μηχανική μάθηση

Η μηχανική μάθηση μπορεί να οριστεί και ως η δυνατότητα ενός συστήματος να μαθαίνει και να βελτιώνει τις επιδόσεις του για συγκεκριμένες διεργασίες ή σύνολο διεργασιών στην πάροδο του χρόνου. Σε αντίθεση με τις στατιστικές τεχνικές οι οποίες επικεντρώνονται στην κατανόηση μιας διεργασίας από το σύνολο των δεδομένων που εξετάζουν, οι τεχνικές μηχανικής μάθησης επικεντρώνονται στην ανάπτυξη ή την βελτίωση του συστήματος βάσει των αποτελεσμάτων που εξήγαγαν από τα δεδομένα σε προηγούμενες χρονικές περιόδους. Αυτό καθιστά ικανό ένα σύστημα που βασίζεται στην μηχανική μάθηση να αλλάζει την στρατηγική εκτέλεσης του βάσει της καινούργιας πληροφορίας που αποκτά κάθε φορά. Αρκετά μοντέλα μηχανικής μάθησης έχουν αναπτυχθεί για την επίλυση προβλημάτων. Τα πιο γνωστά από αυτά είναι τα δίκτυα Bayes, τα Μαρκοβιανά μοντέλα, τα νευρωνικά δίκτυα, η ασαφής λογική (f*uzzy logic*) , οι γενετικοί αλγόριθμοι, η συσταδοποίηση (*clustering*) και οι εύρεση ακραίων τιμών (*outlier detection*).

#### *5.5.3.1 Δίκτυα Bayes (Bayesian Networks)*

Τα δίκτυα Bayes είναι ένα μοντέλο το οποίο κωδικοποιεί τις πιθανολογικές σχέσεις των μεταβλητών που μας ενδιαφέρουν. Αυτή η τεχνική χρησιμοποιείται για ανίχνευση εισβολών σε συνδυασμό με κάποιες στατιστικές τεχνικές. Αυτή η διαδικασία έχει ένα βασικό πλεονέκτημα το οποίο είναι ότι η δυνατότητα να κωδικοποιηθούν αλληλεξαρτήσεις μεταξύ μεταβλητών και να προβλεφθούν γεγονότα [177]. Ωστόσο όπως ειπώθηκε στο [178] η χρήση των δικτυών Bayes έχει ένα μεγάλο μειονέκτημα. Αυτό είναι ότι τα αποτελέσματα τους είναι όμοια με αυτά τα αποτελέσματα που δίνουν τα σύστημα που βασίζονται σε μια τιμή κατωφλίου (*threshold based systems*) αλλά με πολύ μεγαλύτερο υπολογιστικό κόστος. Η χρήση των δικτύων Bayes έχει αποδειχθεί χρήσιμη σε κάποιες καταστάσεις. Παρ' όλα αυτά όμως το γεγονός ότι τα αποτελέσματα βασίζονται σε παραδοχές του τρόπου λειτουργίας εκάστοτε συστήματος. Οπότε σε περίπτωση απόκλισης από την αρχική υπόθεση το σύστημα θα οδηγηθεί σε πολλά λάθη κατά την ανίχνευση.

#### *5.5.3.2 Μαρκοβιανά μοντέλα (Markov models)*

Στα Μαρκοβιανά μοντέλα υπάρχουν δυο προσεγγίσεις. Η μια είναι οι Μαρκοβιανές αλυσίδες [179] οι οποίες είναι ένα σύνολο καταστάσεων οι οποίες είναι αλληλοσυνδεόμενες μεταξύ τους βάσει συγκεκριμένων πιθανοτήτων μετάβασης, οι οποίες καθορίζουν την τοπολογία και τις δυνατότητες του μοντέλου. Κατά την διάρκεια της εκπαίδευσης του συστήματος οι πιθανότητές που σχετίζονται με τις μεταβάσεις υπολογίζονται από την ομαλή συμπεριφορά του συστήματος. Η ανίχνευση των ανωμαλιών πραγματοποιείται συγκρίνοντας την πιθανότητα συσχέτισης η οποία αποκτάται από την παρατήρηση των ακολουθιών και οι οποίες έχουν μια σταθερή τιμή κατωφλίου. Η άλλη προσέγγιση είναι τα Κρυφά Μαρκοβιανά Μοντέλα (*Hidden Markov Models*) [180]. Εδώ το σύστημα θεωρείται ως μια Μαρκοβιανή διαδικασία της οποίας οι καταστάσεις και οι μεταβάσεις είναι κρυφές. Το μόνο που είναι ορατό είναι οι αποκαλούμενες παραγωγές. Τεχνικές που βασίζονται στην θεωρία των Μαρκοβιανών μοντέλων έχουν χρησιμοποιηθεί κατά καιρούς για *Host IDS* [181] και για *Network IDS* [182][183].

#### *5.5.3.3 Νευρωνικά δίκτυα (Neural Networks)*

Τα νευρωνικά δίκτυα προσομοιώνουν κατά κάποιο τρόπο την λειτουργία του ανθρώπινου εγκεφάλου και είναι μια τεχνική η οποία χαρακτηρίζεται από μεγάλη ευελιξία και προσαρμοστικότητα. Στην τεχνική αυτή, το νευρωνικό δίκτυο μαθαίνει πώς να προβλέπει την συμπεριφορά των χρηστών και των διαφορών υποκείμενων του πληροφοριακού συστήματος φτιάχνοντας προφίλ [184]. Η ανίχνευση των εισβολών γίνεται συγκρίνοντας τα προφίλ που ήδη έχει δημιουργήσει με την κίνηση που αναλύει εκείνη την στιγμή

και αν βρει διαφορές τότε θεωρεί ότι υπάρχει εισβολή [185][186]. Το βασικό πλεονέκτημα των νευρωσικών δικτύων είναι η ανοχή που έχουν σε ανακριβή δεδομένα και πληροφορίες που λαμβάνουν σαν είσοδο και η ικανότητα που έχουν να βρίσκουν λύσεις από δεδομένα για τα οποία δεν έχουν προηγούμενη γνώση. Από την άλλη μπορεί να αποτύχουν να βρούνε ικανοποιητική λύση λόγω έλλειψης επαρκών δεδομένων ή επειδή δεν υπάρχει επίκτητη συνάρτηση (*learnable function*). Τέλος τα νευρωνικά δίκτυα είναι μπορεί να είναι αργή και υπολογιστικά ακριβή διαδικασία [187][188][189][190].

#### *5.5.3.4 Τεχνικές ασαφούς λογικής (Fuzzy logic)*

Η ασαφής λογική χρησιμοποιείται στην ανίχνευση εισβολών διότι αρκετές ποιοτικές παράμετροι οι οποίες χρησιμοποιούνται για ανίχνευση εισβολών, όπως για παράδειγμα ο χρόνος χρήσης της κεντρικής μονάδας επεξεργασίας και το χρονικό διάστημα των συνδέσεων, μπορούνε να θεωρηθούν ασαφής μεταβλητές [191][192]. Γενικότερα το η τεχνική της ασάφειας μας διευκολύνει να εξομαλύνουμε την απότομο διαχωρισμό της ομαλής συμπεριφοράς από την μη ομαλή συμπεριφορά. Παρόλου που σαν τεχνική αποδείχτηκε αποδοτική κυρίως σε ανίχνευση επιθέσεων τύπου *port scanning* και *probing*, έχει ένα μεγάλο μειονέκτημα που είναι αυτό της μεγάλης κατανάλωσης υπολογιστικών πόρων [193].

#### *5.5.3.5 Γενετικοί αλγόριθμοι (Genetic Algorithms)*

Οι γενετικοί αλγόριθμοι είναι εμπνευσμένη από την εξελικτική βιολογία. Χρησιμοποιούνται για την εύρεση κατά προσέγγιση λύσεων σε προβλήματα βελτιστοποίησης και αναζήτησης. Έχουν χρησιμοποιηθεί κατά κόρον σε σύστημα ανίχνευσης εισβολών δικτύου [194][195][196] για να διαφοροποιήσουν την ομαλή δικτυακή κίνηση από την μη ομαλή. Το μεγάλο τους πλεονέκτημα είναι η ανθεκτικότητα και η ευελιξία σαν καθολική μέθοδος αναζήτησης. Επιπλέον ένας γενετικός αλγόριθμος ψάχνει να βρει μια λύση εξετάζοντας την από πολλαπλές κατευθύνσεις ενώ δεν υπάρχει προηγουμένη γνώση για την συμπεριφορά του συστήματος και βασίζεται μόνο σε πιθανολογικούς και όχι ντετερμινιστικούς κανόνες.

#### *5.5.3.6 Συσταδοποίηση και ανίχνευση ανωμαλιών*

Η συσταδοποίηση (*clustering*) είναι μια τεχνική για εύρεση μοτίβων σε δεδομένα χωρίς ετικέτες (*unlabeled*) με πολλές διαστάσεις [197][198]. Η συσταδοποίηση έχει προσελκύσει το ενδιαφέρον της ερευνητικής κοινότητας για χρήση σε συστήματα ανίχνευσης εισβολών διότι έχει ένα σημαντικό πλεονέκτημα [20][199][200]. Το πλεονέκτημα αυτό είναι ότι έχει την δυνατότητα να μαθαίνει και ανιχνεύει εισβολές από δεδομένα ελέγχου (*audit data*), χωρίς να απαιτείται από τον διαχειριστή του συστήματος να παρέχει συγκεκριμένες περιγραφές των διαφόρων ειδών και των τύπων επιθέσεων ή εισβολών. Αυτό έχει σαν αποτέλεσμα να χρειάζεται μικρότερος όγκος δεδομένων για τη εκπαίδευση του συστήματος.

Στο [155] οι συγγραφείς κατηγοριοποιούν τις τεχνικές συσταδοποίησης ανίχνευσης ανωμαλιών σε τρεις κατηγορίες βάσει τριών υποθέσεων. Η πρώτη υπόθεση λέει ότι «*τα φυσιολογικά δεδομένα ανήκουν σε μια συστάδα, ενώ τα μη ομαλά δεν ανήκουν σε καμία συστάδα».* Τεχνικές που βασίζονται στην παραπάνω υπόθεση εφαρμόζουν αλγορίθμους συσταδοποίησης και συσταδοποιούν τα δεδομένα. Αν κάποια από τα δεδομένα δεν ανήκουν σε καμία συστάδα τότε αυτά θεωρούνται ως μη ομαλά. Δημοφιλής αλγόριθμοι που λειτουργούν με αυτό τον τρόπο είναι ο DBSCAN [201], ο ROCK [202] και ο SNN [203]. Η δεύτερη υπόθεση λέει ότι *«τα ομαλά δεδομένα βρίσκονται σε κοντινή απόσταση από το κέντρο της συστάδας ενώ τα μη ομαλά βρίσκονται αρκετά μακριά από το κέντρο»*. Οι τεχνικές που υλοποιούν αυτήν την υπόθεση έχουν δυο βήματα. Στο πρώτο βήμα τα δεδομένα ομαδοποιούνται σε συστάδες βάσει ενός αλγορίθμου συσταδοποίησης και στο δεύτερο βήμα μετριέται για κάθε στοιχείο η απόσταση του από το κέντρο της συστάδας. Οι πιο γνωστοί μέθοδοι [204] που βασίζονται σε αυτήν την

υπόθεση είναι οι Χάρτες Αυτό-Οργάνωσης (*Self-Organizing Maps-SOM*), ο αλγόριθμος Κ-μέσων και ο ΕΜ (*Expectation-Maximization*) αλγόριθμος. Η τρίτη υπόθεση λέει το εξής *«τα φυσιολογικά δεδομένα βρίσκονται σε μεγάλες και πυκνές συστάδες ενώ τα μη ομαλά σε μικρές και διάσπαρτες συστάδες»* . Οι τεχνικές που βασίζονται σε αυτήν την υπόθεση κατατάσσουν τα στοιχεία, που βρίσκονται σε συστάδες μικρού μεγέθους ή χαμηλής πυκνότητας σύμφωνα πάντα με κάποιο κατώφλι, ως ανωμαλίες. Αρκετές τέτοιου τύπου τεχνικές έχουν προταθεί στα [205][206][207][208][209]. Η CBLOF [210] είναι αντιπροσωπευτική αυτής της κατηγορίας.

Γενικά υπάρχουν δυο προσεγγίσεις για την ανίχνευση ανωμαλιών με την μέθοδο της συσταδοποίησης. Στην πρώτη προσέγγιση το μοντέλο ανίχνευσης ανωμαλιών εκπαιδεύεται χρησιμοποιώντας δεδομένα χωρίς ετικέτες τα οποία περιλαμβάνουν ομαλή κίνηση και κίνηση με επιθέσεις. Η βασική ιδέα πίσω από αυτό βασίζεται στην υπόθεση ότι οι επιθέσεις και οι εισβολές καταλαμβάνουν ένα μικρό ποσοστό του συνολικού όγκου δεδομένων. Αν αυτή η υπόθεση είναι σωστή τότε οι ανωμαλίες θα μπορούν να ανιχνευθούν βάσει του μεγέθους των συστάδων. Δηλαδή συστάδες με μεγάλο μέγεθος θα περιέχουν την ομαλή κίνηση, ενώ συστάδες μικρότερου μεγέθους θα περιέχουν τις ανωμαλίες οι οποίες είναι και ακραίες τιμές (*outliers*). Στην δεύτερη προσέγγιση το μοντέλο εκπαιδεύεται χρησιμοποιώντας μόνο δεδομένα ομαλής κίνησης και δημιουργεί ένα προφίλ φυσιολογικής συμπεριφοράς.

Τα πλεονεκτήματα της συσταδοποίησης (*clustering*) είναι τα εξής:

- Οι τεχνικές συσταδοποίησης μπορούν να χρησιμοποιηθούν για μη επιβλεπόμενη μάθηση (*unsupervised*).
- Είναι κατάλληλη τεχνική για μεγάλου όγκου σύνολα δεδομένων διότι ομαδοποιούν τα δεδομένα σε συστάδες και είναι ευκολότερος ο εντοπισμός των ανωμαλιών.
- Μπορεί να εφαρμοστεί σε σύνθετα σύνολα δεδομένων εφαρμόζοντας απλά τον αλγόριθμο συσταδοποίησης ο οποίος θα αναλάβει να χειριστεί μόνος του αυτούς του διαφορετικούς τύπους δεδομένων και να δημιουργήσει τις συστάδες.
- Είναι τεχνική υψηλής απόδοσης σε σχέση με τις στατιστικές και γνωσιακές τεχνικές.

Από την άλλη τα βασικά μειονεκτήματα είναι:

- Η απόδοση των τεχνικών συσταδοποίησης βασίζονται στην αποδοτικότητα του αντίστοιχου αλγορίθμου που θα εφαρμοστεί.
- Η παραδοχή ότι οι μεγάλες συστάδες είναι ομαλά δεδομένα και οι μικρές είναι μη ομαλά μπορεί να μην ισχύει πάντα.
- Η δυναμική ανανέωση των προφίλ φυσιολογικής συμπεριφοράς που δημιουργεί είναι μια χρονοβόρα διαδικασία.
- Κάποιες από αυτές τις τεχνικές δεν έχουν δημιουργηθεί με γνώμονα την ανίχνευση ανωμαλιών. Οι ανωμαλίες που εντοπίζουν είναι προϊόν της διαδικασίας συσταδοποίησης και βασίζονται στην υπόθεση ότι όταν κάποιο γεγονός εμφανίζεται σπάνια τότε πιθανόν είναι και ανωμαλία.
- Επειδή όλα τα στοιχεία θα πρέπει να ανήκουν σε μια συστάδα, δεν αποκλείεται μια ανωμαλία να βρίσκεται σε μια μεγάλη συστάδα αντί για μια μικρή με αποτέλεσμα να μην θεωρηθεί επίθεση. Αυτό λόγω της αντίστοιχης υπόθεσης πάνω στην οποία βασίζεται η λειτουργία του αλγορίθμου δεν μπορεί να αποφευχθεί.

- Αρκετές τεχνικές αυτού του είδους είναι αποτελεσματικές μόνο όταν οι ανωμαλίες μπορούν να σχηματίζουν μικρές ή αραιές συστάδες.
- Η επιλογή του κατάλληλου αλγορίθμου με καλή υπολογιστική πολυπλοκότητα παίζει σημαντικό ρόλο για την αποδοτικότητα του συστήματος.

Όπως προαναφέραμε η συσταδοποίηση βοηθά στην εύρεση ανωμαλιών, οι οποίες συνήθως είναι ακραίες τιμές οι οποίες διαφοροποιούνται αρκετά το σύνολο των συστάδων. Με λίγα λόγια από την πλευρά των αλγορίθμων συσταδοποίησης οι ακραίες τιμές είναι αντικείμενα τα οποία δεν μπορούν να κατηγοριοποιηθούν σε συστάδες όμοιου μεγέθους με των υπόλοιπων αντικειμένων. Έτσι μπορούμε να καταλήξουμε στο συμπέρασμα ότι η συσταδοποίηση και η ανίχνευση ακραίων τιμών σχετίζονται μεταξύ τους. Στο [211] αναφέρονται αρκετές τεχνικές εύρεσης ακραίων τιμών. Όταν χρησιμοποιούμε αλγορίθμους εύρεσης ακραίων τιμών η υπόθεση είναι ότι οι ανωμαλίες είναι ασυνήθιστα γεγονότα που συμβαίνουν σε ένα σύστημα. Συστήματα που υλοποιούν αλγορίθμους εύρεσης ακραίων τιμών αναφέρονται στα [212][213]. Τα βασικά πλεονεκτήματα αυτών των μεθόδων είναι ευκολότερο να ανιχνευθούν ακραίες τιμές όταν τα σύνολα των δεδομένων είναι μικρά από ότι όταν είμαι μεγάλα. Σπάνιες επιθέσεις και επιθέσεις τύπου *zero day* μπορούν να ανιχνευθούν αποτελεσματικά με αυτές τις μεθόδους. Από την άλλη τα μειονεκτήματα έχουν να κάνουν με το ότι οι περισσότερες τεχνικές χρησιμοποιούν μεθόδους εύρεσης ακραίων τιμών σε συνδυασμό με μεθόδους συσταδοποίησης, πράγμα που αυξάνει την υπολογιστική πολυπλοκότητα του συστήματος. Τέλος ένα άλλο σημαντικό μειονέκτημα είναι ότι αυτές οι τεχνικές εξαρτώνται σε μεγάλο ποσοστό στο είδος των παραμέτρων (κατηγορηματικές ή συνεχής) τις οποίες καλούνται να επεξεργαστούν.

## 5.7 Σύνοψη

Οι τεχνικές που χρησιμοποιούνται στα παραδοσιακά συστήματα ανίχνευσης εισβολών δεν παρέχουν τον απαραίτητο βαθμό ανιχνευσιμότητας που θα έπρεπε για να προστατέψουν τις πληροφοριακές υποδομές του σήμερα. Όπως προαναφέραμε οι επιθέσεις, οι εισβολές και η πραγματογνωμοσύνη των ατόμων που σχεδιάζουν, υλοποιούν και εξαπολύουν τις επιθέσεις αυτές πλέον υψηλό επίπεδο. Για να αντιμετωπιστούν όλες αυτές οι εκλεπτυσμένες νέου τύπου επιθέσεις πρέπει οι αμυνόμενοι να αναθεωρήσουν τους τρόπους αντιμετώπισης και να εξελίξουν νέες τεχνικές και συστήματα.

Σε αυτό το κεφάλαιο αναφέραμε τα βασικά είδη επιθέσεων και του διαφορετικούς τύπους εισβολέων που πρέπει να αντιμετωπίσει ένα σύστημα ανίχνευσης εισβολών. Επίσης έγινε περιγραφή των τεχνικών που χρησιμοποιούνται σε τέτοιου είδους συστήματα αναφέροντας τα υπέρ και τα κατά κάθε μιας τεχνικής. Αυτή η ενότητα επικεντρώθηκε στην τεχνική ανίχνευσης ανωμαλιών και τις μεθόδους που χρησιμοποιούνται σε αυτήν. Εκτενέστερη αναφορά έγινε στις τεχνικές της μηχανικής μάθησης που δείχνει να μονοπωλεί τον ενδιαφέρον της επιστημονικής κοινότητας λόγω των ευέλικτων και πολλά υποσχόμενων τεχνικών και λύσεων που προσφέρει. Η μηχανική μάθηση προσφέρει λύσεις οι οποίες παρουσιάζουν εξαιρετικά αποτελέσματα στο τομέα της ανίχνευσης εισβολών και όλα αυτά επιβεβαιώνονται από τα διάφορα πειράματα που πραγματοποιήθηκαν και περιγράφονται στα σχετικά επιστημονικά άρθρα που δημοσιεύονται κατά καιρούς. Τέλος καταλήγουμε ότι η τεχνική της ανίχνευσης ανωμαλιών, εκμεταλλευόμενη τις μεθόδους που έχουν αναπτυχθεί για μηχανική μάθηση, είναι η μόνη τεχνική η οποία μπορεί να χειριστεί μεγάλους μεγέθους και διαστάσεων συλλογές δεδομένων. Αυτό καθιστά την ανίχνευση ανωμαλιών μια ιδανική τεχνική για χρήση σε συστήματα ανίχνευσης εισβολών δικτύου (*Network IDS-NIDS*) και μηχανήματων (*Host IDS-HIDS*).

# Κεφάλαιο 6ο

## Σύγχρονες Τεχνικές για Συστήματα Ανίχνευσης Εισβολών

Η έρευνα αυτής της διπλωματικής επικεντρώνεται σε αυτό το κεφάλαιο στις τεχνολογίες οι οποίες θεωρούνται, από την επιστημονική κοινότητα και την ερευνά που διεξήχθη, η αιχμή της τεχνολογίας για την δημιουργία σύγχρονων συστημάτων ανίχνευσης εισβολών (*IDS*). Η έρευνα επικεντρώθηκε στην αναζήτηση τεχνικών οι οποίες θα καταστήσουν τα συστήματα ανίχνευσης εισβολών ικανά να μπορούν να ανταπεξέλθουν γνωστές και άγνωστες εισβολές. Αυτό είναι στις μέρες μας είναι η πρώτη βασική απαίτηση για συστήματα τέτοιου τύπου διότι αυτό είναι το κύριο σημείο στο οποίο υστερούν τα ήδη υπάρχοντα συστήματα ανίχνευσης εισβολών.

Μια δεύτερη απαίτηση είναι ότι αυτά τα συστήματα θα πρέπει να χρειάζονται όσο τον δυνατό λιγότερη ανάδραση και παραμετροποίηση από τον διαχειριστή και γενικότερα τους χρήστες αυτών των συστημάτων. Αυτήν η απαίτηση προκύπτει από το γεγονός ότι τα σημερινά συστήματα ανίχνευσης εισβολών βασίζονται κυρίως στις υπογραφές. Αυτό το πράγμα σημαίνει ότι για να μπορεί να ανιχνεύσει ένα σύστημα τις εισβολές, πρέπει πρώτα αυτές να αναλυθούν από τους ειδικούς και να τους δοθούν υπογραφές. Έπειτα αυτές πρέπει να περαστούν στην βάσει δεδομένων του συστήματος, οπότε σε πιθανή μελλοντική επίθεση το σύστημα θα πρέπει να είναι ικανό να αναγνωρίσει την απειλή από την υπογραφή της. Αυτήν είναι μια χρονοβόρα και κοστοβόρα διαδικασία και συν όλα τα άλλα το σύστημα είναι εκτιθέμενο σε νέες επιθέσεις τύπου *zero day*. Επιπρόσθετα μειώνοντας τον χρόνο ενασχόλησης του χρήστη για συνεχή παραμετροποίηση του συστήματος ανίχνευσης εισβολών του δίνουμε περισσότερο χρόνο για ποιοτικότερη ανάλυση και αντιμετώπιση των διαφόρων συναγερμών ή περιστατικών ασφαλείας που εμφανίζονται κατά καιρούς.

Η τρίτη και τελευταία απαίτηση είναι το σύστημα ανίχνευσης εισβολών να μπορεί να διαχειριστεί μεγάλους όγκους δεδομένων που λαμβάνει από τους αισθητήρες που υπάρχουν διάσπαρτοι στο πληροφοριακό σύστημα, γρήγορα και αποτελεσματικά. Η τελευταία απαίτηση βασίζεται στην φύση της λειτουργίας ενός συστήματος ανίχνευσης εισβολών. Για παράδειγμα ένα σύστημα ανίχνευσης εισβολών που έχει αισθητήρες σε όλο το πληροφοριακό σύστημα (υπολογιστές και συσκευές δικτύου) έχει να διαχειριστεί τεράστιο όγκο δεδομένων. Μόνο και μόνο η διαδικτυακή κίνηση από μόνη της παράγει ένα τεράστιο σύνολο δεδομένων. Για αυτό το συστήματα ανίχνευσης εισβολών θα πρέπει να μπορεί να είναι ικανό να αναλύει μεγάλο όγκο δεδομένων όσο το δυνατόν γρηγορότερα και να εγείρει άμεσα συναγερμούς για τα διάφορα περιστατικά που ανιχνεύει. Όπως είναι γνωστός ο χρόνος είναι πολύτιμος, το ίδιο ισχύει και στα περιστατικά εισβολών. Αν ένα σύστημα ανίχνευσής εισβολών είναι ταχύ και εγείρει συναγερμό γρήγορα τότε δίνει περισσότερο χρόνο στους χρήστες να μπορέσουν να οργανώσουν την άμυνα, να μετριάσουν και να ανακάμψουν από τις επιπτώσεις τις εισβολής (αν αυτή είναι επιτυχής και κατάφερε να επηρεάσει κάποια από τα συστήματα του οργανισμού). Σε αντίθετη περίπτωση αν ανιχνεύσει καθυστερημένα μια εισβολή που ήταν επιτυχής τότε μένει λιγότερος χρόνος στους υπευθύνους να διαχειριστούν την κατάσταση, οι ζημιές που προκαλούνται είναι μεγαλύτερες, ενώ κάποιες φορές μπορεί να είναι και καταστροφικές.

Ένα ακόμα σημαντικό στοιχείο που σχετίζεται με την αποτελεσματικότητα του συστήματος είναι ικανότητα και η ακρίβεια ενός συστήματος στην ανίχνευση εισβολών. Πολλά από αυτά τα συστήματα εγείρουν λανθασμένα συναγερμούς (*false positives*). Οι λανθασμένοι συναγερμοί είναι σχεδόν

αναπόφευκτοι σε ένα τέτοιο σύστημα. Στόχος είναι οι μείωση αυτών στο ελάχιστο, πράγμα που θα μειώσει τον φόρτο εργασίας του χρήστη και θα του εξοικονομήσει πολύτιμο χρόνο.

Για να ικανοποιηθούν οι παραπάνω απαιτήσεις οι έρευνα μας στράφηκε σε άλλους τομείς όμως ο τομέας της εξόρυξης δεδομένων στην οποία παρουσιάζονται τεχνικές για διαχείριση και ανάλυση μεγάλου όγκου πληροφοριών και πολυδιάστατων δεδομένων (*high dimensional data*). Τον τομέα της μηχανικής μάθησης όπου εκεί έχουν αναπτυχθεί διάφορες μέθοδοι οι οποίοι καθιστούν ένα σύστημα ικανό να μαθαίνει από μόνο του και να εξελίσσεται από τα δεδομένα που λαμβάνει κάθε στιγμή. Ιδιαίτερη προσοχή δόθηκε στην μη επιβλεπόμενη μάθηση (*unsupervised learning*) όπου παρουσιάζονται τεχνικές οι οποίες απαιτούν την μικρότερη συμμετοχή του χρήστη κατά την διαδικασία όπου το σύστημα μαθαίνει από τα δεδομένα. Το τρίτο και τελευταίο πεδίο που μελετήθηκε είναι το πεδίο τη ανίχνευσης ανωμαλιών. Αυτό οποίο μελετά μεθόδους και πρακτικές οι οποίες μπορούν μέσα από ένα σύνολο δεδομένων και συγκρίνοντας την συμπεριφορά ενός συστήματος μεταξύ διαφορετικών χρονικών στιγμών, να ανιχνεύσουν μη ομαλές ή ασυνήθιστες συμπεριφορές. Στο δικό μας πρόβλημα αυτό αντιστοιχίζεται με την προσπάθεια ανίχνευσης εισβολών και κυρίως εισβολών τύπου *zero day* για τις οποίες αυτήν στιγμή δεν υπάρχει αποτελεσματικός τρόπος αντιμετώπισης του.

Σε αυτό το κεφάλαιο θα περιγράψουμε την έννοια των πολυδιάστατων δεδομένων με κάποια παραδείγματα, θα μιλήσουμε για την διαδικασία της επιλογής χαρακτηριστικών (*feature selection*) και την πιο εξελιγμένη μορφή αυτής της διαδικασίας την συσταδοποίηση υποχωρών (*subspace clustering*). Επιπρόσθετα θα αναφερθούμε και στην τεχνική της συσταδοποίησης (*clustering*) η οποία μονοπωλεί το ενδιαφέρον των ερευνητών διότι στα πειράματα που διεξήχθησαν από διαφορετικές επιστημονικές ομάδες, τα αποτελέσματα ήταν ενθαρρυντικά για χρήση αυτής τεχνικής για συστήματα που βασίζονται στην μη επιβλεπόμενη (*unsupervised*) μάθηση.

## 6.1 Πολυδιάστατα δεδομένα (*High dimensional data*)

Με τον όρο πολυδιάστατα δεδομένα εννοούμε τα δεδομένα τα οποία έχουν μεγάλο όγκο πληροφοριών δηλαδή πολλές παρατηρήσεις και πολλές μεταβλητές για κάθε παρατήρηση που καταγράφεται [214]. Μια παρατήρηση ή μέτρηση ή καταγραφή μπορεί να είναι οτιδήποτε. Για παράδειγμα μια παρατήρηση μπορεί να είναι μια εικόνα. Αυτή η εικόνα σε ψηφιακή μορφή αποτελείται από ψηφίδες (*pixels*), κάθε ψηφίδα είναι μια διάσταση της εικόνας. Αν σκεφτούμε τώρα ότι πλέον ένα μέσο κινητό τηλέφωνο με κάμερα μπορεί να τραβήξει με φωτογραφία με ανάλυση 10 Megapixel καταλαβαίνουμε αυτόματα ότι οι «διαστάσεις» αυτής της φωτογραφίας είναι μεγάλες, άρα έχουμε μια πολυδιάστατη παρατήρηση. Ως συνέπεια το ο φάκελος του κινητού μας που περιέχει μερικές εκατοντάδες φωτογραφίες (δηλαδή εκατοντάδες παρατηρήσεις) , οι οποίες κάθε μια τους έχουν χιλιάδες διαστάσεις (οι ψηφίδες που αποτελούν την εικόνα) τότε αυτός είναι ένας χώρος πολυδιάστατων δεδομένων. Σε μια βάσει δεδομένων τώρα οι γραμμές είναι οι παρατηρήσεις και οι στήλες είναι οι διαφορές μεταβλητές ή χαρακτηριστικά (*features*) όπως αλλιώς αναφέρονται. Αν αυτήν η βάσει δεδομένων περιέχει κάποιες χιλιάδες γραμμές (παρατηρήσεις) και δεκάδες ή εκατοντάδες στήλες (χαρακτηριστικά) μιλάμε για μια πολυδιάστατη βάσει δεδομένων. Τέτοιου τύπου βάσεις δεδομένων έχουμε σε πολλούς τομείς. Για παράδειγμα στην ιατρική τα δεδομένα που παράγονται από πειράματα μικρό σειρών (*micro array*) για την ανάλυση του DNA, περιλαμβάνουν χιλιάδες γονίδια και εκατοντάδες χαρακτηριστικά (*features*). Χαρακτηριστικά ένα καρκινικό κύτταρο (το οποίο αποτελεί μια παρατήρηση) μπορεί να έχει 6830 χιλιάδες γονίδια (δηλαδή 6830 χαρακτηριστικά) [215]. Επίσης η ανάλυση διαδικτυακής κίνησης παράγει υψηλό όγκο δεδομένων αν σκεφτούμε τα χιλιάδες πακέτα με τις δεκάδες μεταβλητές τους, που

διακινούνται καθημερινά μέσα σε ένα δίκτυο ενός οργανισμού ή μιας εταιρίας. Όπως αναφέρεται στο [216] η εύρεση ανωμαλιών για ανίχνευση εισβολών σε ένα δίκτυο πρέπει να γίνει με τεχνικές οι οποίες μπορούν να διαχειριστούν πολυδιάστατα δεδομένα. Από αυτό καταλαβαίνουμε ότι ένα συστήματα ανίχνευσης εισβολών δικτύου (*NIDS*) θα πρέπει να αναπτυχθεί βάσει τεχνικών ικανών να διαχειριστούν αποτελεσματικά πολυδιάστατα δεδομένα και μεγάλες βάσεις δεδομένων [8]. Για να γίνει καλή διαχείριση αυτού του πολυδιάστατου χώρου δεδομένων έτσι ώστε να μπορέσει να εξαχθεί χρήσιμη πληροφορία, η επιστημονική κοινότητα ανέπτυξε τεχνικές όπως την επιλογή χαρακτηριστικών (*feature selection*) και την συσταδοποίηση υποχωρών (*subspace clustering*).

## 6.2 Επιλογή χαρακτηριστικών (*Feature selection*)

Στον πραγματικό κόσμο τις περισσότερες φορές τα μεγάλα σύνολα δεδομένα που πρέπει να επεξεργαστούμε συνήθως είναι και πολυδιάστατα. Αυτά τα σύνολα απαρτίζονται συνήθως από πολλές χιλιάδες παρατηρήσεις και χαρακτηριστικά. Ωστόσο όσων αφορά τα χαρακτηριστικά που έχει η κάθε παρατήρηση μπορούμε να πούμε ότι δεν είναι όλα χρήσιμα. Ένα τέτοιου είδους σύνολο δεδομένων συνήθως περιέχει πολλά μη συναφή και περιττά χαρακτηριστικά [35]. Ακόμα και οι πιο καλοί αλγόριθμοι δεν μπορούν να διαχειριστούνε αποτελεσματικά σύνολα δεδομένων με πάρα πολλά χαρακτηριστικά εκ των οποίων μερικά από αυτά είναι περιττά ή μη συναφή. Όπως είναι φυσικό αυτά τα περιττά χαρακτηριστικά αλλοιώνουν το αποτέλεσμα των αλγορίθμων (κατηγοριοποίησης , συσταδοποίησης κ.λ.π) που εφαρμόζονται με απώτερο σκοπό την εξαγωγή πληροφοριών και συμπερασμάτων. Επιπλέον οι περισσότεροι αλγόριθμοι δεν μπορούν να διαχειριστούν πολυδιάστατα δεδομένα. Όπως αναφέραμε και προηγουμένως δεδομένα όπως εικόνες, διαδικτυακή κίνηση, αποτελέσματα ιατρικών πειραμάτων και δεδομένα πολυμέσων είναι από την φύση τους πολυδιάστατα. Για τέτοιου είδους ογκώδη δεδομένα εφαρμόζουμε τεχνικές εξόρυξης δεδομένων έτσι ώστε να εξάγουμε πληροφορίες. Μια από αυτές τις τεχνικές που χρησιμοποιούνται στην εξόρυξη δεδομένων είναι και η επιλογή χαρακτηριστικών.

Η επιλογή χαρακτηριστικών έχει ως σκοπό να επιλέξει μόνο τα σημαντικά χαρακτηριστικά και να απαλείψει όσο το δυνατόν γίνεται περιττά και μη συναφή χαρακτηριστικά από ένα σύνολο δεδομένων. Σκοπός είναι η μείωση του συνόλου των χαρακτηριστικών έτσι ώστε να μειωθεί ο υπολογιστικός φόρτος του αλγορίθμου που θα επεξεργαστεί όλα αυτά τα δεδομένα. Επιπλέον θα έχουμε μεγαλύτερη ακρίβεια και ποιότητα πληροφορίας στα αποτελέσματα, πράγμα που τα κάνει ευκολότερο κατανοητά και όπως είναι λογικό πιο εύχρηστα. Για αυτούς τους λόγους η επιλογή χαρακτηριστικών είναι μια κρίσιμη διεργασία στο στάδιο της προ-επεξεργασίας, η οποία πρέπει να γίνει σχολαστικά και λεπτομερώς. Αν αυτήν η διαδικασία είναι επιτυχής αυτό θα φανεί αμέσως κατά την εφαρμογή του αλγορίθμου εξόρυξης δεδομένων και έπειτα στα αποτελέσματα που αυτός θα εξάγει.

Για να γίνει σωστά η διαδικασία της επιλογής χαρακτηριστικών πρέπει να υπάρχει άρτια γνώση του αντικειμένου του δεδομένων και να έχουν ξεκαθαριστεί από την αρχή οι στόχοι που πρέπει να επιτευχθούν έτσι ώστε να φτάσουμε στο επιθυμητό τελικό αποτέλεσμα. Η διαδικασία της επιλογής χαρακτηριστικών μπορεί να γίνει εντελώς χειροκίνητα ή με χρήση κάποιων αυτοματοποιημένων διαδικασιών. Η επιλογή χαρακτηριστικών περιλαμβάνει και μια διαδικασία αναζήτησης στα διάφορα υποσύνολα χαρακτηριστικών με σκοπό την αξιολόγηση του κάθε υποσυνόλου βάσει ενός κριτηρίου [35][217][218][219]. Υπάρχουν τρεις κατηγορίες μεθόδων επιλογής χαρακτηριστικών [220][221] και είναι οι εξής:

- Φίλτρου: Αυτή η μέθοδος πραγματοποιεί την επιλογή χαρακτηριστικών κατά την διάρκεια της προ-επεξεργασίας των δεδομένων χωρίς να έχει ως στόχο την βελτίωση μια συγκεκριμένης

τεχνικής εξόρυξης δεδομένων. Με λίγα λόγια είναι μια τελείως ανεξάρτητη διαδικασία. Για να επιλέξει τα κατάλληλα χαρακτηριστικά πρώτα τα εξετάζει όλα ένα προς ένα και έπειτα εφαρμόζει μια διαδικασία βαθμολόγησης για το καθένα. Για μπορέσει να ξεχωρίσει τα πιο σημαντικά ελέγχει την βαθμολογία και αυτά με την υψηλότερη είναι και τα πιο σημαντικά.

- Περικαλύμματος (*wrappers*): Στην συγκεκριμένη μέθοδο η διαδικασία της επιλογής χαρακτηριστικών περικλείεται-συμπεριλαμβάνεται στον αλγόριθμο που χρησιμοποιούμε για εξόρυξη δεδομένων [222]. Εδώ ένα υποσύνολο χαρακτηριστικών επιλέγεται, η αποτελεσματικότητα του συνόλου αυτού προσδιορίζεται, γίνονται αλλαγές στο αρχικό σύνολο και η αποτελεσματικότητα του νέου συνόλου αξιολογείται ξανά.
- Ενσωματωμένη (*embedded*): Οι ενσωματωμένες μέθοδοι είναι συνδυασμός των δύο προηγούμενων μεθόδων [223].

Αν και η επιλογή χαρακτηριστικών έχει πολλές εφαρμογές και έχει δείξει ενθαρρυντικά αποτελέσματα σε πολυδιάστατα δεδομένα και σε αλγορίθμους συσταδοποίησης (*clustering*), οι οποίοι χρησιμοποιούνται κατά κόρον στην μη επιβλεπόμενη μάθηση, πολλές φορές οι μη συναφής διαστάσεις μπορούν να μπερδέψουν τους αλγόριθμους συσταδοποίησης κρύβοντας συστάδες μέσα σε θορυβώδη δεδομένα. Αυτό σημαίνει ότι οι κλασσικές τεχνικές επιλογής χαρακτηριστικών δεν μας βοηθούν στην μετέπειτα διαδικασία συσταδοποίησης πολυδιάστατων δεδομένων. Για αυτό το λόγο έχουν εξελιχθεί μέθοδοι οι οποίοι είναι επέκταση της επιλογής χαρακτηριστικών και χειρίζονται καλύτερα πολυδιάστατα δεδομένα. Μια τεχνική που είναι δημοφιλής και θεωρείται η επέκταση της επιλογής χαρακτηριστικών είναι η συσταδοποίηση υποχώρου (*subspace clustering*) για την οποία θα μιλήσουμε σε επόμενη παράγραφο.

## 6.3 Συσταδοποίηση (*Clustering*)

Στο κεφάλαιο που μιλήσαμε για μηχανική μάθηση αναφέραμε τον όρο συσταδοποίηση ως μια διαδικασία η οποία συγκεντρώνει και ομαδοποιεί τα αντικείμενα ενός χώρου σε συστάδες (clusters). Τα αντικείμενα κάθε συστάδας έχουν μεταξύ τους μεγάλες ομοιότητες ενώ τα αντικείμενα μεταξύ διαφορετικών συστάδων πρέπει να έχουν διακριτές διαφορές. Πιο απλά αν συγκρίνω ένα αντικείμενο Α με ένα αντικείμενο Β τα οποία βρίσκονται σε μια συστάδα Σ1 πρέπει να είναι βρω μια ομοιότητα s1. Αν πάλι πάρω αυτό το ίδιο αντικείμενο Α και το συγκρίνω με ένα αντικείμενο Γ μιας συστάδας Σ2 τότε οι ομοιότητα s2 του Α με το Γ θα πρέπει να είναι μικρότερη από ότι του Α με το Β.

*Αν ομοιότητα Α και Β = s1 (Α και Β ανήκουν στην Σ1) και η ομοιότητα των Α και Γ = s2 (Α ανήκει στην Σ1 και Γ στην Σ2) τότε πρέπει s1>s2*.

Η συσταδοποίηση είναι η πιο δημοφιλής μέθοδος για μη επιβλεπόμενη μάθηση (unsupervised learning) στην μηχανικής μάθησης. Κάποιες φορές η συσταδοποίηση αναφέρεται και ως μη επιβλεπόμενη μάθηση[33][224].

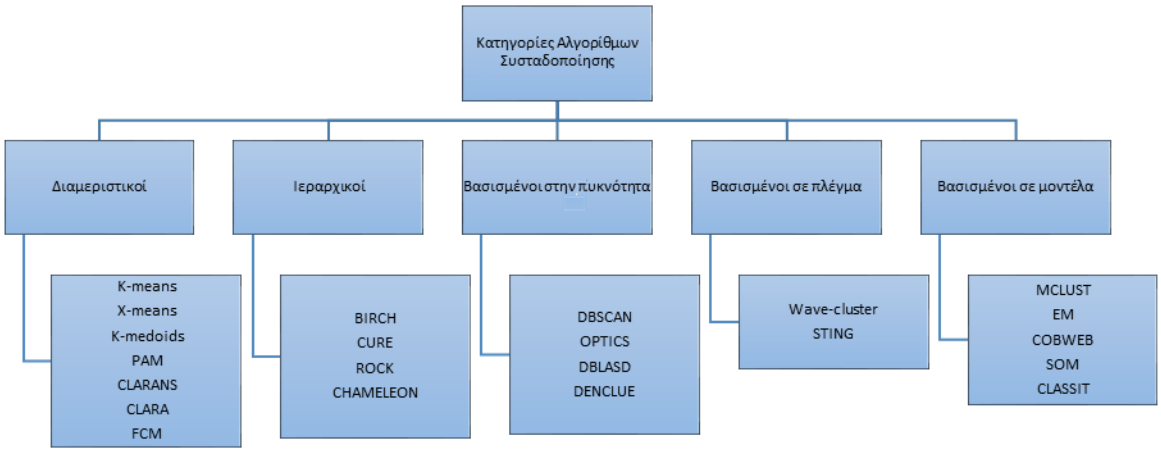
Στην βιβλιογραφία υπάρχει πληθώρα αλγορίθμων που έχουν αναπτυχθεί για συσταδοποίηση. Η κατηγοριοποίηση των μεθόδων που χρησιμοποιούν αυτοί οι αλγόριθμοι είναι μια δύσκολη διαδικασία διότι κάποια μέθοδος μπορεί να δανείζεται χαρακτηριστικά από διαφορετικές κατηγορίες. Ωστόσο όπως παρουσιάζονται και στα [33][98][225][226] οι βασικές κατηγορίες είναι οι εξής:

- Διαμεριστικοί αλγόριθμοι (*Partitioning algorithms*): Σε αυτή την κατηγορία αλγορίθμων η μέθοδος διαχωρισμού χωρίζει τα αντικείμενα του συνόλου των δεδομένων σε πολλές

διαμερίσεις (*partitions*) και κάθε διαμέριση αναπαριστά μια συστάδα (*cluster*). Υπάρχουν δυο βασικές απαιτήσεις που πρέπει να εκπληρώνουν οι συστάδες. Η πρώτη είναι ότι κάθε συστάδα πρέπει να περιέχει τουλάχιστον ένα αντικείμενο και η δεύτερη είναι ότι κάθε αντικείμενο πρέπει να ανήκει σε μια συστάδα. Οι πιο δημοφιλής αλγόριθμοι αυτής της κατηγορίας είναι οι αλγόριθμοι K-means[100], X-means[101] , K-medoids [227], PAM [228],CLARA [229],CLARANS [230] and FCM [231].

- Ιεραρχικοί αλγόριθμοι (*Hierarchy algorithms*): Οι ιεραρχικοί αλγόριθμοι οργανώνονται με ένα ιεραρχικό τρόπο ο οποίος αναπαρίσταται με ένα δενδρόγραμμα. Το δενδρόγραμμα αναπαριστά το σύνολο των δεδομένων ενώ κάθε φύλλο-κόμβος του δενδρογράμματος αναπαριστά κάποιο δεδομένο. Οι ιεραρχικοί μέθοδοι χωρίζονται σε συσσωρευτικής συσταδοποίησης (*agglomerative*) ή αλλιώς από κάτω προς τα επάνω (*bottom up*) και σε διαιρετικής συσταδοποίησης (*divisive*) ή αλλιώς από πάνω προς τα κάτω (*top down*). Στην συσσωρευτική συσταδοποίηση αρχικά κάθε αντικείμενο αναπαριστά μια συστάδα, έπειτα συγχωνεύει τα αντικείμενα ανάλογα με τις μεταξύ τους ομοιότητές σε συστάδες μέχρι να φτάσει σε μια συστάδα που θα τα περιέχει όλα και η οποία θα βρίσκεται στην κορυφή της ιεραρχίας. Η διαιρετική μέθοδος ξεκινά βάζοντας όλα τα αντικείμενα σε μια συστάδα και έπειτα τα διαχωρίζει σε μικρότερες εωσότου κάθε αντικείμενο να αποτελεί μια συστάδα. Είναι το ακριβώς αντίθετο με την συσσωρευτική. Αλγόριθμοι αυτής της κατηγορίας είναι οι BIRCH [232], CURE [233], ROCK [234], CHAMELEON [235].
- Βασισμένοι σε πυκνότητα (*Density-based*): Τα αντικείμενα εδώ χωρίζονται βάσει κριτηρίων όπως είναι η πυκνότητα. Η βασική ιδέα είναι ότι μπορούμε να συνεχίσουμε να αυξάνουμε το μέγεθος μιας συστάδας έως ότου η πυκνότητα (δηλαδή το σύνολο των αντικειμένων) φτάσει σε ένα προκαθορισμένο όριο-κατώφλι. Πιο απλά μπορούμε να πούμε ότι αυτές οι μέθοδοι βασίζονται στην συνεκτικότητα και τις συναρτήσεις πυκνότητας των σημείων των δεδομένων. Οι αλγόριθμοι αυτού του είδους μπορούν να διαχωρίσουν ένα σύνολο αντικειμένων σε πολλαπλές αποκλειστικές συστάδες ή σε μια ιεραρχία συστάδων. Ένα σημαντικό πλεονέκτημα αυτών των αλγορίθμων είναι ότι είναι ικανοί να ανακαλύπτουν ακραίες τιμές (*outlier detection*). Αντιπροσωπευτικοί αλγόριθμοι αυτής της κατηγορίας είναι οι DBSCAN [201], OPTICS [236], DBCLASD [237], DENCLUE [238].
- Βασισμένοι σε πλέγμα (*Grid-based*): Εδώ ο χώρος των αντικειμένων διαιρείται σε ένα πεπερασμένο αριθμό κελιών τα οποία με την σειρά τους σχηματίζουν ένα πλέγμα. Ένα από τα μεγαλύτερα πλεονεκτήματα αυτής της μεθόδου είναι ο γρήγορος χρόνος επεξεργασίας. Αυτό συμβαίνει διότι ο αλγόριθμος διανύει μόνο μια φορά όλο το σύνολο των δεδομένων έτσι ώστε να υπολογίσει τις στατιστικές τιμές για κάθε κελί του πλέγματος. Πιο απλά αυτή η προσέγγιση δεν εξαρτάται από το σύνολο των αντικειμένων αλλά από το σύνολο των κελιών κάθε διάστασης του πλέγματος. Ο Wave-cluster [239] και ο STING [240] είναι αλγόριθμοι που βασίζονται στην ανάλυση πλέγματος.
- Βασισμένοι σε μοντέλα (*Model-based*): Μια τέτοια μέθοδος βελτιστοποιεί την εφαρμογή των δοθέντων δεδομένων με ένα μαθηματικό μοντέλο. Η θεμέλια ιδέα γύρω από αυτές τις μεθόδους είναι ότι τα δεδομένα παράγονται σύμφωνα με μια κατανομή πιθανοτήτων. Υπάρχουν δύο βασικές προσεγγίσεις που επικρατούν εδώ. Η μια είναι η στατιστική προσέγγιση και η άλλη προσέγγιση των νευρωνικών δικτύων. Η στατιστική προσέγγιση χρησιμοποιεί μέτρα πιθανοτήτων για να δημιουργήσει τις συστάδες ενώ τα νευρωνικά δίκτυα κάνουν χρήση κάποιων συνδέσεων μονάδων εισόδου/εξόδου, όπου κάθε σύνδεση έχει ένα «βάρος» (*weight*) το οποίο

σχετίζεται με αυτήν. Τα νευρωνικά δίκτυα έχουν αρκετά πλεονεκτήματα διότι είναι μια κατανεμημένη και παράλληλη αρχιτεκτονική επεξεργασίας πράγμα που τα καθιστά ιδανικά για συσταδοποίηση. Αλγόριθμοι που βασίζονται στην μέθοδο των μοντέλων είναι οι MCLUST [241][242], EM [243], COBWEB [244], SOM [112], CLASSIT [245].




*Εικόνα 4. Κατηγορίες αλγορίθμων συσταδοποίησης.*

## 6.4 Συσταδοποίηση πολυδιάστατων δεδομένων

Τα πολυδιάστατα δεδομένα είναι δεδομένα τα οποία αποτελούνται από ένα μεγάλο αριθμό χαρακτηριστικών (*features*). Ο όρος «κατάρα της διάστασης» (*curse of dimensionality*) αναφέρεται στην αύξηση της πολυπλοκότητας των διάφορων υπολογιστικών προβλημάτων καθώς αυξάνονται οι διαστάσεις των δεδομένων. Η «κατάρα της διάστασης» έχει αποτελέσει το μείζον πρόβλημα των αλγορίθμων συσταδοποίησης διότι έχει καταστήσει τους περισσότερους κλασσικούς αλγορίθμους αναποτελεσματικούς. Καθώς οι διαστάσεις αυξάνονται, ένα πρόβλημα που παρατηρείται είναι ότι χάνεται η ουσιαστική διαφοροποίηση μεταξύ όμοιων και ανόμοιων αντικειμένων. Αυτό οδήγησε την επιστημονική κοινότητα στην μελέτη ανάπτυξης νέων τεχνικών όσων αφορά τις μεθόδους συσταδοποίησης πολυδιάστατων δεδομένων [246].

Οι περισσότεροι παραδοσιακοί αλγόριθμοι συσταδοποίησης είναι ικανοί μόνο για συσταδοποίηση συνόλων δεδομένων με λίγες διαστάσεις. Αυτοί οι αλγόριθμοι βασίζονται σε ομοιότητες ζευγαριών ή ομάδων αντικειμένων έτσι ώστε να τα ομαδοποιήσουν σε συστάδες. Για αυτό τον λόγο αποτυγχάνουν να βρούνε ουσιαστικές συστάδες (*clusters*) σε πολυδιάστατα δεδομένα. Πλέον η εύρεση ομοιοτήτων μεταξύ αντικειμένων χρησιμοποιώντας αποστάσεις (αλγόριθμοι που βασίζονται σε αποστάσεις όπως ο K-means) δεν αποτελεί σωστό κριτήριο, λόγω του μεγάλου-πολυδιάστατου χώρου. Το ίδιο συμβαίνει για τους αλγορίθμους που βασίζονται στην πυκνότητα όπως ο DBSCAN. Επιπλέον καλό είναι να επισημάνουμε ότι οι επιπρόσθετες διαστάσεις εκτός από επιπρόσθετη πληροφορία «κουβαλούν» και επιπρόσθετο θόρυβο. Από αυτήν την άποψη ένας μεγάλος αριθμός διαστάσεων

μπορεί να παρομοιαστεί σαν ένα μεγάλο σύνολο ράδιο-σημάτων, διαφορετικών μεταξύ τους, στα οποία είναι δύσκολο να βρούμε τις διαφορετικές πηγές από τις οποίες αυτά προέρχονται [247].

Η παρουσία μη συναφών χαρακτηριστικών (*irrelevant feature*) ή συσχετίσεων (*correlation*) μεταξύ των υποσυνόλων των χαρακτηριστικών επηρεάζει κατά πολύ την δομή των συστάδων σε ένα πολυδιάστατο χώρο. Το κυρίως πρόβλημα που αντιμετωπίζει εδώ η διαδικασία της συσταδοποίησης είναι ότι διαφορετικά υποσύνολα χαρακτηριστικών διαφορετικών συστάδων, μπορούν να παρουσιάζουν συνάφεια. Επιπρόσθετα διαφορετικές συσχετίσεις μεταξύ χαρακτηριστικών μπορούν να είναι συναφής για διαφορετικές συστάδες. Το φαινόμενο του ότι διαφορετικά χαρακτηριστικά ή συσχετίσεις χαρακτηριστικών μπορεί να είναι συναφή για διαφορετικές συστάδες, αποκαλείται τοπική συνάφεια χαρακτηριστικών *(local feature relevance*) ή τοπική συσχέτιση χαρακτηριστικών (*local feature correlation*). Ένας κοινός τρόπος να ξεπεράσουμε το πρόβλημα χαρακτηριστικών που συσχετίζονται ή είναι συναφή μεταξύ τους σε ένα πολυδιάστατο χώρο δεδομένων είναι να πραγματοποιήσουμε επιλογή χαρακτηριστικών (*feature selection*). Μέθοδοι επιλογής χαρακτηριστικών όπως η Μέθοδος Κύριων Συνιστωσών (*Principal Component Analysis-PCA*) [248], μπορούν να χρησιμοποιηθούν έτσι ώστε αντιστοιχίσουμε τα πρωταρχικό χώρο δεδομένων σε ένα χώρο δεδομένων μικρότερης διάστασης όπου θα μπορέσουμε να συσταδοποιήσουμε τα αντικείμενα του χώρου καλύτερα φτιάχνοντας συστάδες με περισσότερο νόημα και πληροφορία. Δυστυχώς τέτοιες τεχνικές επιλογής χαρακτηριστικών ή μείωσης διαστάσεων (*dimensionality reduction*) δεν μπορούν να εφαρμοστούν στα προβλήματα συσταδοποίησης. Οι τεχνικές επιλογής χαρακτηριστικών και μείωσης διαστάσεων είναι καθολικές τεχνικές (*global techniques*) [249]. Αυτό διότι σαν τεχνικές γενικά υπολογίζουν μόνο ένα υποχώρο (*subspace*) του πρωταρχικού συνόλου δεδομένων στον οποίο έπειτα μπορούμε να εφαρμόσουμε μεθόδους συσταδοποίησης, λαμβάνοντας όμως υπόψη το πλήρες σύνολο των δεδομένων. Σε αντίθεση, το πρόβλημα της τοπικής συνάφειας ή τοπικής συσχέτισης χαρακτηριστικών δηλώνει ότι πολλαπλοί υποχώροι είναι απαραίτητοι, επειδή κάθε συστάδα μπορεί να υπάρξει σε ένα διαφορετικό υποχώρο.

Το συμπέρασμα από όλα τα παραπάνω είναι ότι λόγω του προβλήματος της τοπικής συνάφειας και τοπικής συσχέτισης χαρακτηριστικών συνήθως δεν μπορεί να εφαρμοστεί η καθολική επιλογή χαρακτηριστικών για να ξεπεράσουμε τις δυσκολίες της συσταδοποίησης αντικειμένων σε ένα πολυδιάστατο χώρο δεδομένων. Έτσι λοιπόν αντί για μια καθολική προσέγγιση επιλογής χαρακτηριστικών, μια τοπική προσέγγιση είναι απαραίτητη για την αντιμετώπιση των προβλημάτων τοπικής συνάφειας και τοπικής συσχέτισης χαρακτηριστικών. Μιας και οι κλασσικές μέθοδοι επιλογής χαρακτηριστικών, μείωσης διαστάσεων και συσταδοποίησης δεν είναι αποτελεσματικές, σύγχρονες τεχνικές που ενσωματώνουν την ανάλυση χαρακτηριστικών στην διαδικασία της συσταδοποίησης είναι απαραίτητες.

Αρχικά σημαντικό είναι να αναφέρουμε ότι τα βασικά προβλήματα της συσταδοποίησης σε ένα πολυδιάστατο χώρο είναι δύο. Το πρώτο είναι η αναζήτηση συναφών υποχώρων και το δεύτερο είναι η εύρεση των τελικών συστάδων. Ο χώρος στον οποίο ψάχνουμε για υποχώρους είναι σχεδόν άπειρος. Για να μπορέσουμε να προσεγγίσουμε αυτά τα δύο προβλήματα οι επιστήμονες χρειάστηκε να χρησιμοποιήσουν ευρετικές (*heuristic*) μεθόδους για να αναπτύξουν τις λύσεις τους. Μια από τις πιο διαδεδομένες ευρετικές προσεγγίσεις είναι η προσέγγιση της αναζήτησης υποχώρων (*subspace search*). Η μελέτη αυτής της προσέγγισης οδήγησε την επιστημονική κοινότητα στην ανάπτυξη μεθόδων και αλγορίθμων συσταδοποίησης υποχώρων (*subspace clustering*) οι οποίοι πλέον είναι τα εργαλεία για την επίλυση του προβλήματος της εύρεσης συστάδων σε ένα πολυδιάστατο χώρο [250]. Στην επόμενη

παράγραφο θα αναφερθούμε πιο αναλυτικά στην συσταδοποίηση υποχώρου και στους δημοφιλής αλγορίθμους αυτής της της τεχνικής.

## 6.5 Συσταδοποίηση υποχώρου (*Subspace clustering*)

Η συσταδοποίηση υποχώρου (*subspace clustering*) αναφέρεται στην διαδικασία εύρεσης συστάδων (*clusters*) ομοιών αντικείμενων των οποίων αντικειμένων η ομοιότητα βασίζεται σε ένα συγκεκριμένο υποσύνολο χαρακτηριστικών. Με την συσταδοποίηση υποχώρου και τους αντίστοιχους αλγορίθμους που έχουν αναπτυχθεί ξεπερνάμε τα προβλήματα συσταδοποίησης πολυδιάστατων χώρων των κλασσικών αλγορίθμων συσταδοποίησης. Σε αυτήν την ενότητα θα περιγράψουμε τα είδη των υποχώρων, τις αλγοριθμικές προσεγγίσεις, τις τεχνικές αναζήτησης για την εύρεση κατάλληλων υποχώρων και θα αναφερθούμε στους αλγόριθμους που υπάρχουν.

### 6.5.1 Είδη υποχώρων

Όπως αναφέραμε όταν θέλουμε να συσταδοποιήσουμε αντικείμενα αντιμετωπίζουμε τα φαινόμενα των μη συναφών και των συσχετιζομένων χαρακτηριστικών τα οποία αποτελούνε πρόβλημα κατά την προσπάθεια δημιουργίας συστάδων. Αυτά τα δύο φαινόμενα αναλυθήκαν και οδήγησαν την επιστημονική κοινότητα σε δυο διαφορετικές προσέγγισης και κατά συνέπεια τρία είδη υποχώρων, των οποίων οι ονομασίες έχουν δοθεί βάσει της μεθόδου και της ομάδας των αλγορίθμων συσταδοποίησης που εφαρμόζονται σε κάθε μια αυτές. Οι δύο βασικές προσεγγίσεις είναι οι εξής [251]:

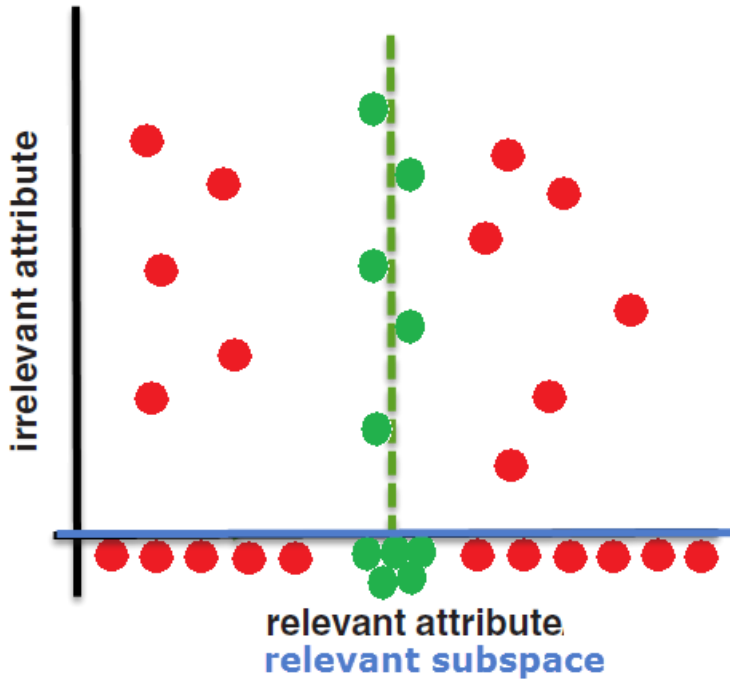
- Υποχώροι παράλληλου άξονα (*axis parallel clustering*): Οι αλγόριθμοί αυτής της κατηγορίας αναφέρονται και ως αλγόριθμοι συσταδοποίησης υποχώρων (*subspace clustering algorithms*) ή αλγόριθμοι προβλεπόμενης συσταδοποίησης (*projected clustering*) [252].
- Υπόχωροι αυθαίρετου προσανατολισμού (*arbitrarily oriented subspaces*): Οι αλγόριθμοι αυτής της κατηγορίας αναφέρονται και ως αλγόριθμοι συσταδοποίησης συσχέτισης (*correlation clustering*) [253].

Εκτός των δυο παραπάνω κατηγοριών προσεγγίσεων υπάρχει και μια τρίτη κατηγορία η οποία περιλαμβάνει αλγορίθμους οι οποίοι αναφέρονται ως αλγόριθμοι βασισμένοι στα μοτίβα (*pattern-based algorithms*). Στην βιβλιογραφία αναφέρονται και ως *biclustering* , *co-clustering* ή *two mode clustering* . Αυτοί οι αλγόριθμοι διευθετούν διαφορετικά είδη υποχώρων και δεν σχετίζονται απευθείας με το φαινόμενο της «κατάρας της διάστασης». Είναι προϊόντα συγκεκριμένων μοντέλων συσταδοποίησης και ακολουθούν μια διαφορετική προσέγγιση από τις παραπάνω που αναφέραμε. Πιο συγκεκριμένα οι αλγόριθμοι αυτής της κατηγορίας ακολουθούν μια υβριδική προσέγγιση μεταξύ των δυο παραπάνω προσεγγίσεων [254][255]. Αλγόριθμοι αυτής της κατηγορίας είναι οι Block Clustering [256], δ-bicluster [257], FLOC [266], p-Cluster [258], MaPIe [259], CoClus [260], OP-Cluster [260].

### 6.5.2 Υποχώροι παράλληλου άξονα (*Axis-Parallel Subspaces*)

Η διαφοροποίηση μεταξύ συναφών (relevant) και μη συναφών (irrelevant) χαρακτηριστικών βασίζεται στην υπόθεση ότι διακύμανση των τιμών των μεταβλητών ενός συναφούς χαρακτηριστικού, σε σχέση με το σύνολο όλων των άλλων σημείων σε μια συστάδα, είναι πιθανόν μικρή σε σύγκριση με την συνολική διακύμανση των τιμών των χαρακτηριστικών. Από την άλλη η διακύμανση των μη συναφών χαρακτηριστικών μιας συστάδας είναι μεγάλη (ή απλά δύσκολο να την διακρίνουμε) αν την συγκρίνουμε με τις τιμές των ίδιων χαρακτηριστικών των άλλων συστάδων. Αν αναπαριστούσαμε γραφικά (βλέπε Εικόνα 5) την παραπάνω περιγραφή θα βλέπαμε ότι στο άξονα των μη συναφών χαρακτηριστικών οι τιμές των μεταβλητών (που τις αναπαριστούμε ως σημεία-κουκκίδες με κόκκινο χρώμα στο γράφημα)

είναι διάσπαρτες (μεγάλη διακύμανση), ενώ στον άξονα των συναφών χαρακτηριστικών (σημεία-κουκίδες με πράσινο χρώμα) είναι πιο πυκνές (μικρή διακύμανση). Όταν επιλέγουμε τα συναφή μόνο χαρακτηριστικά, η συστάδα απεικονίζεται σαν μια συστάδα πλήρους διαστάσεων σε αυτόν υποχώρο (με μπλε γραμμή απεικονίζεται ο άξονας των συναφών χαρακτηριστικών). Στο πλήρη χώρο των διαστάσεων όλα τα σημεία (συμπεριλαμβανομένων συναφών και μη συναφών τιμών των χαρακτηριστικών) της συστάδας σχηματίζουν ένα υπερεπίπεδο (*hyperplane*), το οποίο είναι παράλληλο (πράσινη διακεκομμένη γραμμή) στον άξονα των τιμών των μη συναφών χαρακτηριστικών. Λόγω αυτής της γεωμετρικής απεικόνισης αυτού του είδους η συστάδα ονομάζεται συστάδα παράλληλου υποχώρου [262].




*Εικόνα 5. Συσταδοποίηση παραλλήλου άξονα*

### 6.5.3 Τεχνικές συσταδοποίησης υποχώρου βάσει της μεθόδου αναζήτησης

Οι τεχνικές συσταδοποίησης υποχώρου βάσει της τεχνικής αναζήτησης διακρίνονται σε δύο κατηγορίες. Η πρώτη κατηγορία ακολουθεί την προσέγγιση της από πάνω προς τα κάτω (*top down*) εύρεσης για συσταδοποίηση υποχώρων και ονομάζεται έτσι λόγω της τεχνικής εύρεσης των αλγορίθμων οι οποίοι χρησιμοποιούνται σε αυτήν την προσέγγιση. Η δεύτερη κατηγορία ακολουθεί την προσέγγιση της από κάτω προς τα επάνω (*bottom up*) εύρεσης και πάλι η ονομασία της προέρχεται από τους αντίστοιχους αλγορίθμους που εφαρμόζουν αυτήν την προσέγγιση.

Η από πάνω προς τα κάτω (*top down*) προσέγγιση αρχίζει λαμβάνοντας υπόψη ότι οι συστάδες απαρτίζονται από όλα τα χαρακτηριστικά (διαστάσεις). Σταδιακά αφαιρούνται χαρακτηριστικά και η ποιότητα των συστάδων υπολογίζεται συνεχώς εωσότου η απόδοση τους φτάσει σε ένα συγκεκριμένο ανώτατο όριο. Όπως είναι φυσικό επειδή οι συστάδες αρχικά περιέχουν όλα τα χαρακτηριστικά η επιλογή μια συνάρτησης υπολογισμού αποστάσεων (*distance function*) είναι πολύ σημαντική και

υπάρχει η πιθανότητα μεγάλου υπολογιστικού φόρτου. Πιο συγκεκριμένα η προσέγγιση ξεκινά δημιουργώντας συστάδες που η κάθε μια τους περιέχει όλο το σύνολο των χαρακτηριστικών που υπάρχουν στο χώρο. Αυτές οι συστάδες είναι επίσης ίσης διάστασης. Για κάθε διάσταση υπάρχει και μια τιμή η οποία αποκαλείται βάρος και δείχνει την σπουδαιότητα-αξία κάθε διάστασης. Για κάθε διάσταση κάθε συστάδας ορίζεται και μια συγκεκριμένη τιμή βάρους. Στην συνέχεια γίνονται επαναλήψεις και σταδιακά αφαιρούνται διαστάσεις. Έπειτα ξανά υπολογίζονται οι τιμές βάρους οι οποίες χρησιμοποιούνται για την δημιουργία καινούργιων συστάδων. Αυτήν η επαναληπτική διαδικασία απαιτεί πολλαπλές επαναλήψεις υπολογιστικά ακριβών αλγορίθμων συσταδοποίησης σε ένα πλήρες σύνολο διαστάσεων. Οι αλγόριθμοι αυτής της τεχνικής δημιουργούν συστάδες οι οποίες είναι μέρος του ευρύτερου συνόλου δεδομένων πράγμα που σημαίνει ότι κάθε δείγμα που χρησιμοποιούν αντιστοιχίζεται σε μια μόνο συστάδα. Πολλές από αυτές τις τεχνικές χρησιμοποιούν την τεχνική της δειγματοληψίας για να βελτιώσουν την απόδοση τους. Όπως καταλαβαίνουμε η απόδοση τους εξαρτάται άμεσα από την ποιότητα και το μέγεθος του δείγματος τα οποία παίζουν καθοριστικό ρόλο στο τελικό αποτέλεσμα [263]. Δημοφιλής αλγόριθμοι που χρησιμοποιούν αυτήν την τεχνική είναι οι PROCLUS [252], ORCLUS [264] , FINDIT [265], δ-Clusters [266], COSA [267].

Η από κάτω προς τα πάνω (*bottom up*) προσέγγιση εκμεταλλεύεται την ιδιότητα της κλειστότητας προς τα κάτω (*downward closure property*) για να μειώσει το χώρο αναζήτησης χρησιμοποιώντας μια προσέγγιση που μοιάζει με την APRIORI. Οι αλγόριθμοι που ακολουθούν αυτήν την προσέγγιση χρησιμοποιούν δομές δεδομένων όπως κελιά, πλέγματα ή μονάδες. Κατά την έναρξη της διαδικασίας, πρώτα δημιουργούν ένα ιστόγραμμα για κάθε διάσταση και επιλέγουν αυτές τις διαστάσεις των οποίων η πυκνότητα είναι πάνω από ένα συγκεκριμένο κατώφλι. Η ιδιότητα της κλειστότητας προς τα κάτω δηλώνει ότι αν υπάρχουν πυκνές μονάδες σε ένα σύνολο *κ* διαστάσεων τότε υπάρχουν και πυκνές μονάδες σε όλες τις *κ-1* προβολές των διαστάσεων. Έπειτα υποψήφιοι υποχώροι σε δυο διαστάσεις μπορούν να δημιουργηθούν, χρησιμοποιώντας μόνο εκείνες τις διατάσεις οι οποίες περιέχουν πυκνές μονάδες. Αυτό μειώνει κατά πολύ τον χώρο αναζήτησης του αλγορίθμου. Ο αλγόριθμους τρέχει εωσότου δεν υπάρξουν άλλες πυκνές μονάδες. Έπειτα αυτές οι πυκνές μονάδες χρησιμοποιούνται για την δημιουργία πυκνών συστάδων. Η φύση των αλγορίθμων αυτής της προσέγγισης οδηγεί σε επικαλυπτόμενες συστάδες όπου κάθε παρατήρηση μπορεί να βρίσκεται σε πολλές συστάδες ή σε καμία συστάδα. Η εξαγωγή καλών αποτελεσμάτων εξαρτάται κυρίως από την σωστή παραμετροποίηση του μεγέθους των πλεγμάτων και τις παραμέτρους κατωφλίου που καθορίζουν την πυκνότητα. Η παραμετροποίηση των τελευταίων μπορεί να είναι αρκετά δύσκολη διαδικασία ειδικά αν σκεφτούμε ότι πρέπει να γίνει για το σύνολο των διαστάσεων του συνόλου δεδομένων [268][269]. Δημοφιλής αλγόριθμοι που εφαρμόζουν αυτήν την προσέγγιση είναι οι CLIQUE [103],ENCLUS [270], MAFIA [271], CBF [272], DOC [273], CLTree [274].

### 6.5.4 Αλγόριθμοι συσταδοποίησης βάσει της διαδικασίας εύρεσης υποχώρων

Λόγω του μεγάλου χώρου αναζήτησης, όλοι οι αλγόριθμοι της κατηγορίας παράλληλου άξονα, για να βρούνε τους κατάλληλους υποχώρους βασίζονται σε κάποιες υποθέσεις. Στην βιβλιογραφία υπάρχουν τέσσερις κατηγορίες αλγορίθμων σύμφωνα με την υπόθεση στην οποία βασίζονται [250]. Αυτές οι κατηγορίες αλγορίθμων είναι οι εξής:

- Προβαλόμενης συσταδοποίησης (*projected clustering*): Αυτοί οι αλγόριθμοι στοχεύουν στο να βρούνε μια μοναδική εκχώρηση κάθε σημείου του συνόλου δεδομένων, σε ακριβώς μια συστάδα υποχώρου. Γενικά προσπαθούνε να βρούνε μια προβολή ενός συνόλου σημείων τα οποία μπορούνε να συσταδοποιηθούν με τον καλύτερο δυνατό τρόπο [275][276].

Δημοφιλέστερη αλγόριθμοι σε αυτήν την κατηγορία είναι οι PROCLUS [252], SSPC [277] ,PreDeCon [278].

- «Απαλής» προβαλόμενης συσταδοποίησης *("soft" projected clustering*): Μερικοί αλγόριθμοι προβαλόμενης συσταδοποίησης προϋποθέτουν από πριν των αριθμό των συστάδων που θα δημιουργηθούν, έτσι ώστε να προσδιοριστεί μια συνάρτηση η οποία θα είναι βέλτιστη και θα έχει ως σκοπό να βρει τον ιδανικό σετ συστάδων. Αυτήν η προσέγγιση δανείζεται την φιλοσοφία της συσταδοποίησης του αλγορίθμου k-means [283]. Εδώ στα διαφορετικά χαρακτηριστικά δίνονται διαφορετικές τιμές βάρους αλλά όλα τα χαρακτηριστικά συμβάλουν στην τελική δημιουργία των συστάδων [284][285][286][288]. Αλγόριθμοι αυτής της κατηγορίας είναι ο COSA [267] και ο LAC [287].
- Συσταδοποίησης υποχώρων (*subspace clustering*): Σε αυτήν την κατηγορία αλγορίθμων στόχος είναι η εύρεση όλων τον υποχώρων όπου συστάδες μπορούν να εξευρεθούν. Οι αλγόριθμοι εδώ έχουν ως κύριο σκοπό την εύρεση συστάδων σε όλους τους υποχώρους. Γνωστοί αλγόριθμοι αυτής της κατηγορίας είναι ο CLIQUE [103], ENCLUS [270], MAFIA [271], SUBCLU [278], DUSC [289], nCluster [290].
- Υβριδικοί αλγόριθμοι (h*ybrid algorithms*): Η τέταρτη και τελευταία κατηγορία αλγορίθμων έχει ως στόχο την εύρεση συστάδων οι οποίες πιθανόν να επικαλύπτονται. Οι αλγόριθμοι αυτής της κατηγορίας δεν στοχεύουν στην εύρεση όλων των συστάδων σε όλους τους υποχώρους. Κάποιοι υβριδικοί αλγόριθμοι υπολογίζουν μόνο ενδιαφέροντες υποχώρους αντί για συστάδες υποχώρων. Τέλος οι αναφερόμενοι υπόχωροι μπορούν να εξαχθούν εφαρμόζοντας αλγορίθμους πλήρους διάστασης σε αυτές τις προβολές. Αλγόριθμοι που ανήκουν σε αυτήν την κατηγορία είναι οι COSA [267], DOC [273], MINECLUS [279][280], DiSH [291], HARP [292], SCHISM [293], FIRES [294], P3C [281][282][283].

Τέλος πρέπει να επισημάνουμε ότι αλγόριθμοι της προσέγγισης παράλληλου άξονα λειτουργούν βάσει των δύο τεχνικών αναζήτησης (της από πάνω προς τα κάτω και της από κάτω προς τα πάνω αναζήτησης) που αναφέραμε πιο πάνω. Οι αλγόριθμοι της συσταδοποίησης υποχώρου ακολουθούν όλοι την τεχνική της από κάτω προς τα πάνω αναζήτησης (*bottom up*), ενώ οι αλγόριθμοι της προβαλλόμενης συσταδοποίησης ακολουθούν στην πλειοψηφία τους την από πάνω προς τα κάτω (*top down*) τεχνική, πλην του P3C ο οποίος εφαρμόζει την από κάτω προς τα πάνω. Οι υβριδικοί αλγόριθμοι από την άλλη όπως οι DiSH, FIRES, P3C, SCHISM χρησιμοποιούν την από κάτω προς τα πάνω αναζήτηση ενώ οι DOC ,MINECLUS, COSA, HARP εφαρμόζουν την από πάνω προς τα κάτω (βλέπε **Error! Reference source not found.** ).

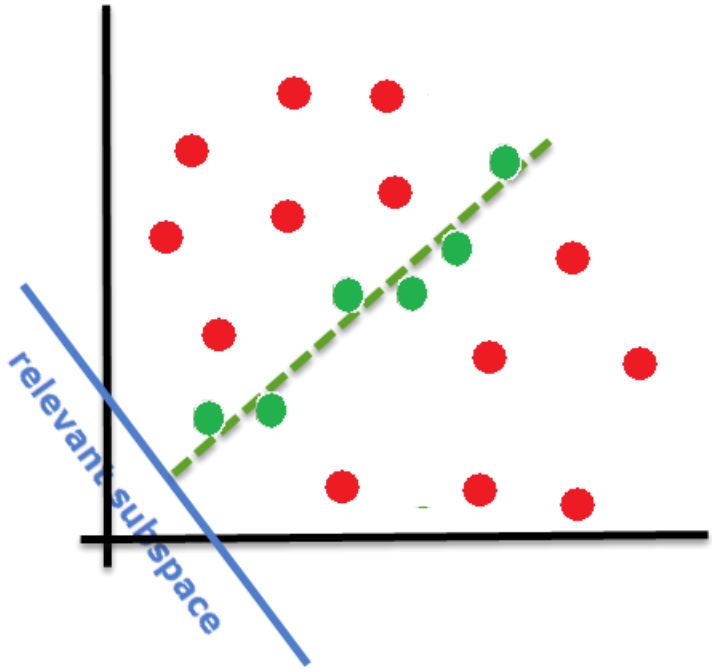
| Κατηγορία αλγορίθμων | Προσέγγιση βάσει της μεθόδου αναζήτησης | |
|---|---|---|
| | Από κάτω προς τα πάνω (bottom-up) | Από πάνω προς τα κάτω (top-down) |
| Συσταδοποίηση Υποχώρου | CLIQUE<br>ENCLUS<br>MAFIA<br>SUBCLU<br>nCluster | |
| Υβριδικοί Αλγόριθμοι | FIRES<br>P3C | DOC<br>MINECLUS |

| | | |
|---|---|---|
| | DiSH<br>SCHISM | COSA<br>HARP |
| Προβαλόμενης Συσταδοποίησης | | PROCLUS<br>SSPC<br>PreDeCon |

*Πίνακας ταξινόμησης αλγορίθμων βάσει της μεθόδου αναζήτησης*

### 6.5.5 Υποχώροι αυθαίρετου προσανατολισμού (*arbitrarily oriented subspaces*)

Αν δύο χαρακτηριστικά α και β συσχετίζονται για ένα σύνολο σημείων, τα σημεία αυτά θα είναι διάσπαρτα κατά μήκος ενός υπέρ-επιπέδου το οποίο ορίζεται από κάποια γραμμική εξάρτηση μεταξύ των δύο παραπάνω χαρακτηριστικών και η οποία γραμμική εξάρτηση αντιστοιχεί στην παραπάνω συσχέτιση. Ο υποχώρους ο οποίος είναι ορθογώνιος σε αυτό το υπερεπίπεδο είναι ένας υποχώρος όπου τα σημεία συσταδοποιούνται με μεγάλη πυκνότητα, ανεξάρτητα από την διακύμανση του συνδυασμού των τιμών α και β. Αυτός ο υποχώρος είναι αυθαίρετα προσανατολισμένος (βλέπε Εικόνα 6). Αυτήν η περίπτωση είναι πιο γενική σε σχέση με την περίπτωση του παράλληλου άξονα [251].




*Εικόνα 6. Υποχώροι αυθαίρετου προσανατολισμού.*

### 6.5.6 Βασικές τεχνικές και αλγόριθμοι

Από αυτά που αναφέρθηκαν πιο πάνω το συμπέρασμα που προκύπτει είναι ότι οι γραμμικές εξαρτήσεις έχουν ως αποτέλεσμα τις ισχυρές γραμμικές συσχετίσεις μεταξύ χαρακτηριστικών. Οπότε αυτό το είδος συσταδοποίησης μπορεί να ονομαστεί και ως συσταδοποίηση συσχετίσεων (*correlation clustering*). Οι βασικές τεχνικές για την εύρεση υποχώρων αυθαίρετου προσανατολισμού είναι η Μέθοδος Κύριων Συνιστωσών *(Principal Component Analysis-PCA*) και η μετατροπή του Hough [295].

Η πλειοψηφία των αλγορίθμων της συσταδοποίησης συσχετίσεων βασίζεται στην μέθοδο PCA. Η εφαρμογή της PCA μεθόδου επηρεάζει κατά πολύ την χρονική πολυπλοκότητα αυτών των αλγορίθμων η οποία συνήθως είναι ο αριθμός των διαστάσεων υψωμένος εις τον κύβο. Ο πρώτος αλγόριθμος που εφάρμοσε την τεχνική αυτήν ήταν ο ORLCUS [296]. Έπειτα ακολουθήσαν αλγόριθμοι όπως ο 4C [297], COPAC [298], ERiC [299].

Στην μέθοδο μετατροπής Hough βασίζονται αλγόριθμοι όπως CASH [300]. Αυτήν η μέθοδος προτάθηκε από τον Achtert κ.α. στα [300][301]. Η βασική ιδέα αυτής της μεθόδου είναι ή αντιστοίχιση ενός σημείου του χώρου δεδομένων σε μια συνάρτηση του αποκαλουμένου χώρου παραμέτρων. Η μέθοδος αυτήν είχε αναπτυχθεί αρχικά για ανάλυση πολυδιάστατων δεδομένων έχοντας ως περιεχόμενο εικόνες. Ο αρχικός σχεδιασμός αυτής της μεθόδου ήταν η αντιστοίχιση σημείων ενός δυσδιάστατου χώρου (αποκαλούμενος και ως χώρος εικόνας) Ευκλείδειων συντεταγμένων (π.χ. ψηφίδες μια εικόνας), σε ένα χώρο παραμέτρων. Άλλα παραδείγματα παρόμοιων τεχνικών είναι οι τεχνικές φίλτρου εικόνας όπως χρησιμοποιεί ο αλγόριθμος MrCC [302]. Επίσης τεχνικές τυχαίας δειγματοληψίας έχουν χρησιμοποιηθεί όπως αυτήν στον αλγόριθμο RANSAC [303]. Τέλος πρέπει να αναφέρουμε ότι το μεγαλύτερο μειονέκτημα των αλγορίθμων συσχέτισης είναι η υπόθεση που κάνουν ότι τα σημεία της συστάδας βρίσκονται κοντά μεταξύ τους (δηλαδή μια συστάδα εμφανίζεται ως ένα σύνολο πυκνά κατανεμημένων σημείων) μέσα στον Ευκλείδειο χώρο.

## 6.6 Σύνοψη

Συνοψίζοντας σε αυτό το κεφάλαιο αναφερθήκαμε σε θέματα όπως οι χώροι πολυδιάστατων δεδομένων και τα προβλήματα ανάλυσης και εξαγωγής πληροφοριών από αυτούς. Υπενθυμίζουμε ότι τα περισσότερα προβλήματα της καθημερινότητας μας και οι συλλογές δεδομένων που συλλέγονται από τα πληροφορικά συστήματα είναι έχουν πολυδιάστατη φύση. Για αυτό το λόγο η επιστημονική κοινότητα προσπαθεί συνεχώς να αναπτύξει μεθόδους και τεχνικές ανάλυσης και εξαγωγής χρήσιμης πληροφορίας από αυτά τα μεγάλα πολυδιάστατα σύνολα δεδομένων. Τέτοιες τεχνικές είναι η επιλογή χαρακτηριστικών η οποία προσπαθεί να μειώσει τις διαστάσεις και να επικεντρωθεί μόνο στα χρήσιμα χαρακτηριστικά και τις τιμές τους, πριν προχωρήσουμε περαιτέρω την ανάλυση με την τεχνική της συσταδοποίησης. Η συσταδοποίηση από την μεριά της είναι μια αποτελεσματική προσέγγιση η οποία αποτελείται από αρκετές μεθόδους για ομαδοποίηση και για κατηγοριοποίηση αντικείμενων ενός μεγάλου χώρου δεδομένων. Επιπρόσθετα εξελίξεις των παραδοσιακών τεχνικών ήρθαν να δώσουν λύσεις στα προβλήματα της επιλογής χαρακτηριστικών και συσταδοποίησης πολυδιάστατων χώρων. Τέτοιες εξελιγμένες τεχνικές είναι η συσταδοποίηση υποχώρων η οποία περιλαμβάνει ποικιλία τεχνικών και αλγορίθμων οι οποίοι είναι ικανοί να αντιμετωπίσουν τα προβλήματα των πολυδιάστατων χώρων δεδομένων.

Η συσταδοποίηση υποχώρων έχει εφαρμογή σε πολλούς τομείς και χρησιμοποιείται κατά κόρον από τον κλάδο της ιατρικής και της βιολογίας για την ανάλυση πολυδιάστατων συνόλων δεδομένων που περιέχουν πληροφορίες σχετικές με το DNA και τα γονίδια. Για αυτό τον λόγο είναι μια τεχνική η οποία αρχίζει να μελετάται πολύ και από τους επιστήμονες του χώρου της πληροφορικής οι οποίοι ειδικεύονται στην ανάπτυξη συστημάτων προστασίας υπολογιστών, όπως είναι τα συστήματα ανίχνευσης εισβολών [14]. Τα συστήματα ανίχνευσης εισβολών αντιμετωπίζουν πλήθος πολυδιάστατων δεδομένων τα οποία δεν μπορούν να αντιμετωπιστούν με τις παραδοσιακές μεθόδους με τις οποίες είναι αυτήν την στιγμή υλοποιημένα τα περισσότερα συστήματα. Οπότε η ανάπτυξη συστημάτων με τις σύγχρονες τεχνικές οι

οποίες μπορούν να διαχειριστούν μεγάλους όγκους πολυδιάστατων δεδομένων είναι πλέον επιτακτική ανάγκη.

# Κεφάλαιο 7ο

## Συμπεράσματα

Αυτήν η εργασία είχε ως σκοπό την διερεύνηση τεχνικών οι οποίες μπορούν θα μπορέσουν να χρησιμοποιηθούν για την ανάπτυξη συστημάτων ανίχνευσής εισβολών νέας γενιάς. Αυτά τα συστήματα νέας γενιάς όπως τονίσαμε στα προηγούμενα κεφάλαια θα πρέπει να λειτουργούν σχεδόν αυτόματα και να χρειάζονται την λιγότερη δυνατή ανάδραση από τους χρήστες που είναι υπεύθυνοι για την ορθή λειτουργία και συντήρηση αυτών των συστημάτων. Επιπλέον θα πρέπει να διαχειρίζονται και να αναλύουν τεράστιο όγκο δεδομένων όσο το δυνατό γρηγορότερα έτσι ώστε να μπορούν να ανιχνεύουν τις επιθέσεις έγκαιρα πριν αυτές προκαλέσουν ζημιά στο πληροφοριακό σύστημα. Πέραν αυτού, ακόμα μια από τις προϋποθέσεις της ερευνάς μας ήταν ότι θα πρέπει αυτά τα συστήματα να έχουν μηχανισμούς με τους οποίους θα μπορούν να ανιχνεύουν πρωτοεμφανιζόμενες επιθέσεις τύπου *zero day,* για τις οποίες δεν υπάρχει προηγούμενη γνώση ή πιο απλά δεν είναι ευρέως γνωστή η ύπαρξη τους. Με βάσει τα παραπάνω η ερευνά μας είχε ως κεντρικούς άξονες την μελέτη τεχνικών οι οποίες θα μπορέσουν να δώσουν στο σύστημα την ικανότητα να δρα από μόνο του σε κάποιες περιστάσεις και θα μπορούν να αναλύσουν ταχύτατα μεγάλους όγκος δεδομένων.

Η έρευνα μας ξεκίνησε αρχικά με την μελέτη τεχνικών εξόρυξης δεδομένων έτσι ώστε να κατανοήσουμε τον τρόπο με τον οποίο εξάγεται γνώση από μεγάλους όγκους δεδομένων. Από την εξόρυξη δεδομένων οδηγηθήκαμε στην μηχανική μάθηση η οποία προσφέρει λύσεις αυτοματοποίησης σε πολλούς κλάδους και χρησιμοποιείται πλέον στις περισσότερες εφαρμογές που αναλύουν δεδομένα με σκοπό την εξαγωγή χρήσιμης πληροφορίας. Πιο συγκεκριμένα επικεντρωθήκαμε στην μη επιβλεπόμενη μηχανική μάθηση έτσι ώστε να κατανοήσουμε τους τρόπους με τους οποίους ένα σύστημα μπορεί να μαθαίνει από μόνο του αναλύοντας και επεξεργάζονται τα δεδομένα που έχει δέχεται σαν είσοδο. Με αυτήν την τεχνική ένα σύστημα μπορεί να καταστεί σχετικά ευφυές έτσι ώστε να μπορεί να λαμβάνει αποφάσεις σε κάποιο βαθμό, χωρίς να χρειάζεται απαραίτητα κάθε φορά ανθρώπινη παρέμβαση. Αρκετές μέθοδοι και αλγόριθμοι της μη επιβλεπόμενης μηχανικής μάθησης, όπως οι αλγόριθμοι συσταδοποίησης έχουν παρουσιάσει ενθαρρυντικά αποτελέσματα κατά την εφαρμογή τους για ανίχνευση εισβολών αναλύοντας διαδικτυακή κίνηση. Αυτό την καθιστά μια από τις πιο δημοφιλής μεθόδους σε συστήματα ανίχνευσης εισβολών δικτύου (*NIDS*) πράγμα που αποδεικνύεται από τις διάφορες επιστημονικές δημοσιεύσεις γύρω από αυτό το θέμα.

Η τεχνική της συσταδοποίησης και οι αλγόριθμοί της έχουν μελετηθεί εκτενέστερα και για την χρήση τους στην ανίχνευση ανωμαλιών όπως αναφέραμε στο κεφάλαιο 5. Τα αποτελέσματα πειραμάτων που δημοσιεύονται κατά καιρούς σε επιστημονικά άρθρα [14][304][305][306][307] δείχνουν η ανίχνευση ανωμαλιών είναι εφικτή με την χρήση των μεθόδων της συσταδοποίησης. Ωστόσο το μεγαλύτερο πρόβλημα όπως το επισημάναμε και στο κεφάλαιο 6 είναι ότι ένα σύστημα ανίχνευσης εισβολών έχει να αντιμετωπίσει εκτός από μεγάλα σύνολα δεδομένων και πολυδιάστατα δεδομένα. Για την αντιμετώπιση του προβλήματος των πολυδιάστατων δεδομένων τεχνικές όπως η συσταδοποίηση υποχώρων έρχονται να δώσουν λύση. Ωστόσο περαιτέρω έρευνα χρειάζεται για την βελτιστοποίηση των μεθόδων και των αλγορίθμων συσταδοποίησης υποχώρου έτσι ώστε να αυξηθεί η αποτελεσματικότητά τους και να μπορέσουν στο μέλλον να εφαρμοστούν πρακτικά για την ανάπτυξη συστημάτων όπως αυτά της ανίχνευσης εισβολών (*IDS*), τα οποία θα είναι αξιόπιστα και θα έχουν ικανοποιητικά ποσοστά ανιχνευσιμότητας σε επιθέσεις τύπου *zero day*.

# Αναφορές


1. http://data.worldbank.org/indicator/IT.NET.USER.ZS?end=2015&start=1990&view=chart Last accessed: 19/04/2017
2. http://searchsecurity.techtarget.com/report/Recent-ransomware-attacks-Data-shows-50-growth-in-2016 Last accessed: 19/04/2017
3. http://www.express.co.uk/news/world/766154/banks-warning-computer-hackers-finances-attacked-Kaspersky Last accessed: 19/04/2017
4. Denning, Dorothy E. *"An intrusion-detection model."* IEEE Transactions on software engineering 2 (1987): 222-232.
5. Wu, Shelly Xiaonan, and Wolfgang Banzhaf. *"The use of computational intelligence in intrusion detection systems: A review."* Applied Soft Computing 10.1 (2010): 1-35.
6. Khan, Javed Akhtar, and Nitesh Jain. *"A Survey on Intrusion Detection Systems and Classification Techniques."* (2016).
7. De Boer, Pieter, and Martin Pels. *"Host-based intrusion detection systems."* Amsterdam University (2005).
8. Amoli, Payam Vahdani, and Timo Hamalainen. *"A real time unsupervised NIDS for detecting unknown and encrypted network attacks in high speed network."* Measurements and Networking Proceedings (M&N), 2013 IEEE International Workshop on. IEEE, 2013.
9. Akamai's State of the Internet, Q4 Report (2016). https://www.akamai.com/StateOfTheInternet Last accessed: 14/2/2017
10. Garcia-Teodoro, Pedro, et al. *"Anomaly-based network intrusion detection: Techniques, systems and challenges."* computers & security 28.1 (2009): 18-28.
11. Jyothsna, V., VV Rama Prasad, and K. Munivara Prasad. *"A review of anomaly based intrusion detection systems."* International Journal of Computer Applications 28.7 (2011): 26-35.
12. Amoli, Payam Vahdani, and Timo Hamalainen. *"A real time unsupervised NIDS for detecting unknown and encrypted network attacks in high speed network."* Measurements and Networking Proceedings (M&N), 2013 IEEE International Workshop on. IEEE, 2013.
13. Bhuyan, Monowar H., D. K. Bhattacharyya, and Jugal K. Kalita. *"An effective unsupervised network anomaly detection method."* Proceedings of the International Conference on Advances in Computing, Communications and Informatics. ACM, 2012.
14. Casas, Pedro, Johan Mazel, and Philippe Owezarski. *"Unsupervised network intrusion detection systems: Detecting the unknown without knowledge."* Computer Communications 35.7 (2012): 772-783.
15. Gupta, Megha. *"Hybrid Intrusion Detection System: Technology and Development."* International Journal of Computer Applications 115.9 (2015).
16. Zhang, Yu-Fang, Zhong-Yang Xiong, and Xiu-Qiong Wang. *"Distributed intrusion detection based on clustering."* Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on. Vol. 4. IEEE, 2005.
17. Vasilomanolakis, Emmanouil, et al. *"Taxonomy and survey of collaborative intrusion detection."* ACM Computing Surveys (CSUR) 47.4 (2015): 55.
18. da Silva, Ana Paula R., et al. *"Decentralized intrusion detection in wireless sensor networks."* Proceedings of the 1st ACM international workshop on Quality of service & security in wireless and mobile networks. ACM, 2005.

19. Huang, Yi-an, and Wenke Lee. *"A cooperative intrusion detection system for ad hoc networks."* Proceedings of the 1st ACM workshop on Security of ad hoc and sensor networks. ACM, 2003.
20. Portnoy, Leonid, Eleazar Eskin, and Sal Stolfo. *"Intrusion detection with unlabeled data using clustering."* In Proceedings of ACM CSS Workshop on Data Mining Applied to Security (DMSA-2001. 2001.
21. Garcia-Teodoro, Pedro, et al. *"Anomaly-based network intrusion detection: Techniques, systems and challenges."* computers & security 28.1 (2009): 18-28.
22. Liao, Hung-Jen, et al. *"Intrusion detection system: A comprehensive review."* Journal of Network and Computer Applications 36.1 (2013): 16-24.
23. Scarfone, Karen, and Peter Mell. *"Guide to intrusion detection and prevention systems (idps)."* NIST special publication 800.2007 (2007): 94.
24. Joshi, Manish. *"Classification, clustering and intrusion detection system."* International Journal of Engineering Research and Applications (IHERA) 2.2 (2012): 961-964.
25. Bhuyan, Monowar H., Dhruba Kumar Bhattacharyya, and Jugal K. Kalita. *"Network anomaly detection: methods, systems and tools."* IEEE communications surveys & tutorials 16.1 (2014): 303-336.
26. Chebrolu, Srilatha, Ajith Abraham, and Johnson P. Thomas. *"Feature deduction and ensemble design of intrusion detection systems."* Computers & security 24.4 (2005): 295-307.
27. Douligeris, Christos, and Aikaterini Mitrokotsa. *"DDoS attacks and defense mechanisms: classification and state-of-the-art."* Computer Networks 44.5 (2004): 643-666.
28. Lee, Wenke, and Salvatore J. Stolfo. *"A framework for constructing features and models for intrusion detection systems."* ACM transactions on Information and system security (TiSSEC) 3.4 (2000): 227-261.
29. Paliwal, Swati, and Ravindra Gupta. *"Denial-of-service, probing & remote to user (R2L) attack detection using genetic algorithm."* International Journal of Computer Applications 60.19 (2012): 57-62.
30. Debar, Hervé, Marc Dacier, and Andreas Wespi. *"Towards a taxonomy of intrusion-detection systems."* Computer Networks 31.8 (1999): 805-822.
31. Kumar, Koushal, and Jaspreet Singh Batth. *"Network Intrusion Detection with Feature Selection Techniques using Machine-Learning Algorithms."* International Journal of Computer Applications 150.12 (2016).
32. Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. *"From data mining to knowledge discovery in databases."* AI magazine 17.3 (1996): 37.
33. Han, Jiawei, Jian Pei, and Micheline Kamber. "*Data mining: concepts and techniques*". Elsevier, 2011.
34. Osmar R. Zaïane, "*Chapter I: Introduction to Data Mining CMPUT 690 Principles of Knowledge Discovery in Databases*", 1999.
35. Liu, Huan, and Hiroshi Motoda. "*Feature selection for knowledge discovery and data mining*". Vol. 454. Springer Science & Business Media, 2012.
36. Liu, Huan, and Rudy Setiono. *"Chi2: Feature selection and discretization of numeric attributes."* Tools with artificial intelligence, 1995. proceedings., seventh international conference on. IEEE, 1995.
37. Cios, Krzysztof J., Witold Pedrycz, and Roman W. Swiniarski. "*Data mining methods for knowledge discovery*." Vol. 458. Springer Science & Business Media, 2012.

38. Cox, David R. *"Regression models and life-tables."* Breakthroughs in statistics. Springer New York, 1992. 527-541.
39. Long, J. Scott, and Jeremy Freese. "*Regression models for categorical dependent variables using Stata.*" Stata press, 2006.
40. Draper, Norman Richard, Harry Smith, and Elizabeth Pownell. "*Applied regression analysis*." Vol. 3. New York: Wiley, 1966.
41. Agrawal, Rakesh, et al. *"Fast Discovery of Association Rules."* Advances in knowledge discovery and data mining 12.1 (1996): 307-328.
42. Zembowicz, Robert, and Jan M. Żytkow. *"From contingency tables to various forms of knowledge in databases."* Advances in knowledge discovery and data mining. American Association for Artificial Intelligence, 1996.
43. Getoor, Lise, and Christopher P. Diehl. *"Link mining: a survey."* ACM SIGKDD Explorations Newsletter 7.2 (2005): 3-12.
44. Srivastava, Jaideep, et al. *"Web usage mining: Discovery and applications of usage patterns from web data."* Acm Sigkdd Explorations Newsletter 1.2 (2000): 12-23.
45. Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas. *"Supervised machine learning: A review of classification techniques."* (2007): 3-24.
46. Halkidi, Maria, Yannis Batistakis, and Michalis Vazirgiannis. *"On clustering validation techniques."* Journal of intelligent information systems 17.2-3 (2001): 107-145.
47. Berkhin, Pavel. *"A survey of clustering data mining techniques."* Grouping multidimensional data. Springer Berlin Heidelberg, 2006. 25-71.
48. Hodge, Victoria J., and Jim Austin. *"A survey of outlier detection methodologies."* Artificial intelligence review 22.2 (2004): 85-126.
49. He, Zengyou, Xiaofei Xu, and Shengchun Deng. *"Discovering cluster-based local outliers."* Pattern Recognition Letters 24.9 (2003): 1641-1650.
50. Kantardzic, Mehmed. "*Data mining: concepts, models, methods, and algorithms*." John Wiley & Sons, 2011.
51. Markou, Markos, and Sameer Singh. *"Novelty detection: a review—part 1: statistical approaches."* Signal processing 83.12 (2003): 2481-2497.
52. Chen, Ming-Syan, Jiawei Han, and Philip S. Yu. *"Data mining: an overview from a database perspective."* IEEE Transactions on Knowledge and data Engineering 8.6 (1996): 866-883.
53. Alpaydin, Ethem. "*Introduction to machine learning*". MIT press, 2014.
54. Witten, Ian H., et al. "*Data Mining: Practical machine learning tools and techniques*." Morgan Kaufmann, 2016.
55. Wu, Xindong, et al. *"Top 10 algorithms in data mining."* Knowledge and information systems 14.1 (2008): 1-37.
56. Keim, Daniel A. *"Information visualization and visual data mining."* IEEE transactions on Visualization and Computer Graphics 8.1 (2002): 1-8.
57. Fayyad, Usama M., Andreas Wierse, and Georges G. Grinstein. "*Information visualization in data mining and knowledge discovery*." Morgan Kaufmann, 2002.
58. Agrawal, Rakesh, and Ramakrishnan Srikant. *"Privacy-preserving data mining."* ACM Sigmod Record. Vol. 29. No. 2. ACM, 2000.
59. Tavani, Herman T. *"Informational privacy, data mining, and the internet."* Ethics and Information Technology 1.2 (1999): 137-145.

60. Clifton, Chris, et al. *"Tools for privacy preserving distributed data mining."* ACM Sigkdd Explorations Newsletter 4.2 (2002): 28-34.
61. Agrawal, Dakshi, and Charu C. Aggarwal. *"On the design and quantification of privacy preserving data mining algorithms."* Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. ACM, 2001.
62. Patcha, Animesh, and Jung-Min Park. *"An overview of anomaly detection techniques: Existing solutions and latest technological trends."* Computer networks 51.12 (2007): 3448-3470.
63. Lee, Wenke, Salvatore J. Stolfo, and Kui W. Mok. *"Adaptive intrusion detection: A data mining approach."* Artificial Intelligence Review 14.6 (2000): 533-567.
64. Lee, Wenke, Salvatore J. Stolfo, and Kui W. Mok. *"A data mining framework for building intrusion detection models."* Security and Privacy, 1999. Proceedings of the 1999 IEEE Symposium on. IEEE, 1999.
65. Lappas, Theodoros, and Konstantinos Pelechrinis. *"Data mining techniques for (network) intrusion detection systems."* Department of Computer Science and Engineering UC Riverside, Riverside CA 92521 (2007).
66. Bama, S. Sathya, MS Irfan Ahmed, and A. Saravanan. *"Network intrusion detection using clustering: a data mining approach."* International Journal of Computer Applications 30.4 (2011).
67. Lee, Wenke, et al. *"Real time data mining-based intrusion detection."* DARPA Information Survivability Conference & Exposition II, 2001. DISCEX'01. Proceedings. Vol. 1. IEEE, 2001.
68. Pietraszek, Tadeusz, and Axel Tanner. *"Data mining and machine learning—towards reducing false positives in intrusion detection."* Information security technical report 10.3 (2005): 169-183.
69. Marr, Marr. *"A Short History of Machine Learning - Every Manager Should Read."* Forbes. Retrieved 27 Feb 2016.
70. Samuel, Arthur L. *"Some studies in machine learning using the game of checkers."* IBM Journal of research and development 3.3 (1959): 210-229.
71. Russell, Stuart, Peter Norvig, and Artificial Intelligence. *"A modern approach."* Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs 25 (1995): 27.
72. Marsland, Stephen. "*Machine learning: an algorithmic perspective.*" CRC press, 2015
73. Maimon, Oded, and Lior Rokach. *"Introduction to supervised methods."* Data Mining and Knowledge Discovery Handbook. Springer US, 2005. 149-164.
74. Nicolas, Patrick R. "*Scala for Machine Learning.*" Packt Publishing Ltd, 2015.
75. Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar. "*Foundations of machine learning.*" MIT press, 2012.
76. Caruana, Rich, and Alexandru Niculescu-Mizil. *"An empirical comparison of supervised learning algorithms."* Proceedings of the 23rd international conference on Machine learning. ACM, 2006.
77. Yan, Xin, and Xiaogang Su. "*Linear regression analysis: theory and computing.*" World Scientific, 2009.
78. Utgoff, Paul E. *"Incremental induction of decision trees."* Machine learning 4.2 (1989): 161-186.
79. Quinlan, J. Ross. *"Induction of decision trees."* Machine learning 1.1 (1986): 81-106.
80. Gunn, Steve R. *"Support vector machines for classification and regression."* ISIS technical report 14 (1998): 85-86.
81. Meyer, David, and FH Technikum Wien. *"Support vector machines."* The Interface to libsvm in package e1071 (2015).

82. Weinberger, Kilian Q., and Lawrence K. Saul. *"Distance metric learning for large margin nearest neighbor classification."* Journal of Machine Learning Research 10.Feb (2009): 207-244.
83. Suguna, N., and K. Thanushkodi. *"An improved k-nearest neighbor classification using genetic algorithm."* International Journal of Computer Science Issues 7.2 (2010): 18-21.
84. Navot, Amir, et al. "*Nearest neighbor based feature selection for regression and its application to neural activity."* Advances in Neural Information Processing Systems 18 (2006): 995.
85. Rish, Irina. *"An empirical study of the naive Bayes classifier."* IJCAI 2001 workshop on empirical methods in artificial intelligence. Vol. 3. No. 22. IBM New York, 2001.
86. Murphy, Kevin P. *"Naive bayes classifiers."* University of British Columbia (2006).
87. Ho, Tin Kam. *"Random decision forests."* Document Analysis and Recognition, 1995, Proceedings of the Third International Conference on. Vol. 1. IEEE, 1995.
88. Ho, Tin Kam. "*The random subspace method for constructing decision forests."* IEEE transactions on pattern analysis and machine intelligence 20.8 (1998): 832-844.
89. Breiman, Leo. *"Random forests."* Machine learning 45.1 (2001): 5-32.
90. Carpenter, Gail A., et al. "*Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps."* IEEE Transactions on neural networks 3.5 (1992): 698-713.
91. Ojha, Varun Kumar, Ajith Abraham, and Václav Snášel. *"Metaheuristic design of feedforward neural networks: A review of two decades of research."* Engineering Applications of Artificial Intelligence 60 (2017): 97-116.
92. Zhang, Guoqiang Peter. *"Neural networks for classification: a survey."* IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 30.4 (2000): 451-462.
93. Rosenblatt, Frank. x. (1961). "*Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms."*
94. Collobert, Ronan, and Samy Bengio. "*Links between perceptrons, MLPs and SVMs."* Proceedings of the twenty-first international conference on Machine learning. ACM, 2004.
95. Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. "*Unsupervised learning. The elements of statistical learning."* Springer New York, 2009. 485-585.
96. Atiya, Amir F. *"An unsupervised learning technique for artificial neural networks."* Neural Networks 3.6 (1990): 707-711.
97. Barbara, Daniel, and Sushil Jajodia. *"Applications of data mining in computer security."* Vol. 6. Springer Science & Business Media, 2002.
98. Xu, Rui, and Donald Wunsch. *"Survey of clustering algorithms."* IEEE Transactions on neural networks 16.3 (2005): 645-678.
99. Berkhin, Pavel. *"A survey of clustering data mining techniques."* Grouping multidimensional data. Springer Berlin Heidelberg, 2006. 25-7.
100. Jain, Anil K. *"Data clustering: 50 years beyond K-means."* Pattern recognition letters 31.8 (2010): 651-666.
101. Pelleg, Dan, and Andrew W. Moore. "*X-means: Extending K-means with Efficient Estimation of the Number of Clusters."* ICML. Vol. 1. 2000.
102. Zhao, Ying, and George Karypis. *"Evaluation of hierarchical clustering algorithms for document datasets."* Proceedings of the eleventh international conference on Information and knowledge management. ACM, 2002.

103. Agrawal, Rakesh, et al. "*Automatic subspace clustering of high dimensional data for data mining applications."* Vol. 27. No. 2. ACM, 1998.
104. McLachlan, Geoffrey J., and Kaye E. Basford. *"Mixture models: Inference and applications to clustering."* Vol. 84. Marcel Dekker, 1988.
105. Figueiredo, Mario A. T., and Anil K. Jain. *"Unsupervised learning of finite mixture models."* IEEE Transactions on pattern analysis and machine intelligence 24.3 (2002): 381-396.
106. Chandola, Varun, Arindam Banerjee, and Vipin Kumar. *"Anomaly detection: A survey."* ACM computing surveys (CSUR) 41.3 (2009): 15.
107. Lin, Chin-Teng, and CS George Lee. *"Neural fuzzy systems."* PTR Prentice Hall (1996).
108. Radford, Alec, Luke Metz, and Soumith Chintala. "*Unsupervised representation learning with deep convolutional generative adversarial networks."* 2015.
109. Roweis, Sam T., and Lawrence K. Saul. *"Nonlinear dimensionality reduction by locally linear embedding"* .science 290.5500 (2000): 2323-2326.
110. Fodor, Imola K. *"A survey of dimension reduction techniques."* Center for Applied Scientific Computing, Lawrence Livermore National Laboratory 9 (2002): 1-18.
111. Kohonen, Teuvo. *"The self-organizing map."* Neurocomputing 21.1 (1998): 1-6.
112. Cottrell, Marie, Jean-Claude Fort, and Gilles Pagès. *"Theoretical aspects of the SOM algorithm"* .Neurocomputing 21.1 (1998): 119-138.
113. Ayodele, Taiwo Oladipupo. *"Types of machine learning algorithms."* INTECH Open Access Publisher, 2010.
114. Zhu, Xiaojin. *"Semi-supervised learning literature survey."* 2005.
115. Zhu, Xiaojin, and Andrew B. Goldberg. *"Introduction to semi-supervised learning"* .Synthesis lectures on artificial intelligence and machine learning 3.1 (2009): 1-130.
116. Belkin, Mikhail, and Partha Niyogi. *"Semi-supervised learning on Riemannian manifolds."* Machine learning 56.1-3 (2004): 209-239.
117. Rosenberg, Chuck, Martial Hebert, and Henry Schneiderman. *"Semi-supervised self-training of object detection models."* 2005.
118. Fazakis, Nikos, et al. *"Self-trained LMT for semisupervised learning."* Computational intelligence and neuroscience 2016 (2016): 10.
119. Culp, Mark, and George Michailidis. *"An iterative algorithm for extending learners to a semi-supervised setting."* Journal of Computational and Graphical Statistics 17.3 (2008): 545-571.
120. Blum, Avrim, and Tom Mitchell. *"Combining labeled and unlabeled data with co-training."* Proceedings of the eleventh annual conference on Computational learning theory. ACM, 1998.
121. Zhou, Zhi-Hua, and Ming Li. *"Semi-Supervised Regression with Co-Training."* IJCAI. Vol. 5. 2005.
122. Chapelle, Olivier, Vikas Sindhwani, and Sathiya S. Keerthi. *"Optimization techniques for semi-supervised support vector machines."* Journal of Machine Learning Research 9.Feb (2008): 203-233.
123. Chapelle, Olivier, Bernhard Scholkopf, and Alexander Zien. *"Semi-supervised learning."* IEEE Transactions on Neural Networks 20.3 (2009): 542-542.
124. Sutton, Richard S., and Andrew G. Barto. *"Reinforcement learning: An introduction."* Vol. 1. No. 1. Cambridge: MIT press, 1998.
125. Busoniu, Lucian, et al. *"Reinforcement learning and dynamic programming using function approximators."* Vol. 39. CRC press, 2010.

126. Williams, Ronald J. *"A class of gradient-estimating algorithms for reinforcement learning in neural networks."* Proceedings of the IEEE First International Conference on Neural Networks. Vol. 2. 1987.
127. Szepesvári, Csaba. *"Algorithms for reinforcement learning."* Synthesis lectures on artificial intelligence and machine learning 4.1 (2010): 1-103.
128. Zhang, Jun, et al. *"Evolutionary computation meets machine learning: A survey."* IEEE Computational Intelligence Magazine 6.4 (2011): 68-75.
129. Eiben, Agoston E., and James E. Smith. *"Introduction to evolutionary computing."* Vol. 53. Heidelberg: springer, 2003.
130. Mitchell, Melanie. *"An introduction to genetic algorithms."* MIT press, 1998.
131. http://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/ Last accessed: 8/3/2107
132. http://www.bbc.co.uk/timelines/zc6fbk7 Last accessed 14/3/2017
133. http://www.independent.co.uk/life-style/gadgets-and-tech/news/cyber-criminal-iphone-android-smartphone-fitbit-smart-tv-national-cyber-security-centre-natinoal-a7627976.html Last accessed 14/3/2017
134. Anderson, James P. *"Computer security threat monitoring and surveillance."* Vol. 17. Technical report, James P. Anderson Company, Fort Washington, Pennsylvania, 1980.
135. Kayacik, H. Günes, A. Nur Zincir-Heywood, and Malcolm I. Heywood. *"Selecting features for intrusion detection: A feature relevance analysis on KDD 99 intrusion detection datasets"* .Proceedings of the third annual conference on privacy, security and trust. 2005.
136. Ghorbani, Ali A., Wei Lu, and Mahbod Tavallaee. *"Network intrusion detection and prevention: concepts and techniques."* Vol. 47. Springer Science & Business Media, 2009.
137. Hoque, Nazrul, et al. *"Network attacks: Taxonomy, tools and systems."* Journal of Network and Computer Applications 40 (2014): 307-324.
138. Gaudin, Sharon. *"Case study of insider sabotage: the Tim Lloyd/Omega case."* Computer Security Journal 16.3 (2000): 1-9.
139. Axelsson, Stefan. "*Intrusion detection systems: A survey and taxonomy*." Vol. 99. Technical report, 2000.
140. Sundaram, Aurobindo. *"An introduction to intrusion detection."* Crossroads 2.4 (1996): 3-7.
141. Gupta, Megha. *"Hybrid Intrusion Detection System: Technology and Development."* International Journal of Computer Applications 115.9 (2015).
142. Aydın, M. Ali, A. Halim Zaim, and K. Gökhan Ceylan. "*A hybrid intrusion detection system design for computer network security."* Computers & Electrical Engineering 35.3 (2009): 517-526.
143. Thottan, Marina, and Chuanyi Ji. *"Anomaly detection in IP networks."* IEEE Transactions on signal processing 51.8 (2003): 2191-2204.
144. Thottan, Marina, and C. Ji. *"Using network fault predictions to enable IP traffic management."* Journal of Network and Systems Management 9.3 (2001): 327-346.
145. Maxion, Roy A., and Frank E. Feather. *"A case study of ethernet anomalies in a distributed computing environment."* IEEE transactions on Reliability 39.4 (1990): 433-443.
146. Vigna, Giovanni, and Richard A. Kemmerer. *"NetSTAT: A network-based intrusion detection approach."* Computer Security Applications Conference, 1998. Proceedings. 14th Annual. IEEE, 1998.

147. Yang, Jiahai, et al. *"CARDS: A distributed system for detecting coordinated attacks."* Information Security for Global Information Infrastructures. Springer US, 2000. 171-180.
148. Wang, Haining, Danlu Zhang, and Kang G. Shin. *"Detecting SYN flooding attacks."* INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE. Vol. 3. IEEE, 2002.
149. Savage, Stefan, et al. *"Practical network support for IP traceback."* ACM SIGCOMM Computer Communication Review. Vol. 30. No. 4. ACM, 2000.
150. Axelsson, Stefan. *"Research in intrusion-detection systems: A survey."* Vol. 120. Technical report 98–17. Department of Computer Engineering, Chalmers University of Technology, 1998.
151. Estevez-Tapiador, Juan M., Pedro Garcia-Teodoro, and Jesus E. Diaz-Verdejo. *"Anomaly detection methods in wired networks: a survey and taxonomy."* Computer Communications 27.16 (2004): 1569-1584.
152. Kohavi, Ron. "*A study of cross-validation and bootstrap for accuracy estimation and model selection."* Ijcai. Vol. 14. No. 2. 1995.
153. Kumar, Sandeep, and Eugene H. Spafford. *"An application of pattern matching in intrusion detection."* 1994.
154. Lazarevic, Aleksandar, Vipin Kumar, and Jaideep Srivastava. *"Intrusion detection: A survey."* Managing Cyber Threats. Springer US, 2005. 19-78.
155. Chandola, Varun, Arindam Banerjee, and Vipin Kumar. *"Anomaly detection: A survey."* ACM computing surveys (CSUR) 41.3 (2009): 15.
156. Patcha, Animesh, and Jung-Min Park. *"An overview of anomaly detection techniques: Existing solutions and latest technological trends."* Computer networks 51.12 (2007): 3448-3470.
157. Garcia-Teodoro, Pedro, et al. *"Anomaly-based network intrusion detection: Techniques, systems and challenges."* Computers & security 28.1 (2009): 18-28.
158. Jyothsna, V., VV Rama Prasad, and K. Munivara Prasad. *"A review of anomaly based intrusion detection systems."* International Journal of Computer Applications 28.7 (2011): 26-35.
159. Bhuyan, Monowar H., Dhruba Kumar Bhattacharyya, and Jugal K. Kalita. *"Network anomaly detection: methods, systems and tools."* IEEE communications surveys & tutorials 16.1 (2014): 303-336.
160. Anscombe, Frank J. *"Rejection of outliers."* Technometrics 2.2 (1960): 123-146.
161. Motulsky, H. "*Intuitive Biostatistics: Choosing a statistical test."* Oxford University Press (1995): Chapter 37.
162. Van Doorn, Erik A. *"Quasi-stationary distributions and convergence to quasi-stationarity of birth-death processes."* Advances in Applied Probability 23.04 (1991): 683-700.
163. Denning, Dorothy E., and Peter G. Neumann. *"Requirements and model for IDES—a real-time intrusion detection expert system."* Document A005, SRI International 333 (1985).
164. Ye, Nong, et al. *"Multivariate statistical analysis of audit trails for host-based intrusion detection."* IEEE Transactions on computers 51.7 (2002): 810-820.
165. Barnet, V. *"The ordering of multivariate data (with discussion)."* Journal of the Royal Statistics Society, Series A 139 (1976): 318-354.
166. Pires, A., and Carla Santos-Pereira. *"Using clustering and robust estimators to detect outliers in multivariate data."* Proceedings of the International Conference on Robust Statistics. 2005.

167. Valley, Fraser, et al. "*DETECTING HACKERS (ANALYZING NETWORK TRAFFIC) BY POISSON MODEL MEASURE."* http://www2.ensc.sfu.ca/people/grad/pwangf/IPSW_report.pdf Last accessed: 19/03/17
168. Abraham, Bovas, and George EP Box. *"Bayesian analysis of some outlier problems in time series."* Biometrika (1979): 229-236.
169. Abraham, Bovas, and Alice Chuang. *"Outlier detection and time series modeling."* Technometrics 31.2 (1989): 241-248.
170. Beckman, Richard J., and R. Dennis Cook. *"Outlier……. s."* Technometrics 25.2 (1983): 119-149.
171. Sekar, R., et al. *"Specification-based anomaly detection: a new approach for detecting network intrusions."* Proceedings of the 9th ACM conference on Computer and communications security. ACM, 2002.
172. Estevez-Tapiador, Juan M., Pedro Garcia-Teodoro, and Jesus E. Diaz-Verdejo. *"Stochastic protocol modeling for anomaly based network intrusion detection"* .Information Assurance, 2003. IWIAS 2003. Proceedings. First IEEE International Workshop on. IEEE, 2003.
173. Noel, Steven, Duminda Wijesekera, and Charles Youman. *"Modern intrusion detection, data mining, and degrees of attack guilt."* Applications of data mining in computer security. Springer US, 2002. 1-31.
174. Lunt, Teresa F., Ann Tamaru, and F. Gillham. "A real-time intrusion-detection expert system (IDES)." SRI International. Computer Science Laboratory, 1992.
175. Lee, Wenke, and Salvatore J. Stolfo. *"Data Mining Approaches for Intrusion Detection."* Usenix security. 1998.
176. Lee, Wenke, Salvatore J. Stolfo, and Kui W. Mok. "Adaptive intrusion detection: A data mining approach." Artificial Intelligence Review 14.6 (2000): 533-567.
177. Heckerman, David. *"A tutorial on learning with bayesian networks."* Microsoft Research. 1995.
178. Kruegel, Christopher, et al. *"Bayesian event classification for intrusion detection."* Computer Security Applications Conference, 2003. Proceedings. 19th Annual. IEEE, 2003.
179. Ye, Nong, Yebin Zhang, and Connie M. Borror. *"Robustness of the Markov-chain model for cyber-attack detection."* IEEE Transactions on Reliability 53.1 (2004): 116-123.
180. Hu, Jiankun, et al. *"A simple and efficient hidden Markov model scheme for host-based anomaly intrusion detection."* IEEE network 23.1 (2009): 42-47.
181. Yeung, Dit-Yan, and Yuxin Ding. *"Host-based intrusion detection using dynamic and static behavioral models."* Pattern recognition 36.1 (2003): 229-243.
182. Mahoney, Matthew V., and Philip K. Chan. "*Learning nonstationary models of normal network traffic for detecting novel attacks."* Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2002.
183. Estevez-Tapiador, Juan M., Pedro García-Teodoro, and Jesús E. Díaz-Verdejo. *"Detection of web-based attacks through Markovian protocol parsing."* Computers and Communications, 2005. ISCC 2005. Proceedings. 10th IEEE Symposium on. IEEE, 2005.
184. Fox, Kevin L., et al. *"A neural network approach towards intrusion detection."* (1990): 124-134.
185. Debar, Herve, Monique Becker, and Didier Siboni. *"A neural network component for an intrusion detection system."* Research in Security and Privacy, 1992. Proceedings., 1992 IEEE Computer Society Symposium on. IEEE, 1992.
186. Cansian, Adriano M., et al. *"Network intrusion detection using neural networks."* Proc. Int. Conf. on Computational Intelligence and Multimedia Applications. 1997.

187. Ghosh, Anup K., Christoph Michael, and Michael Schatz. *"A real-time intrusion detection system based on learning program behavior."* International Workshop on Recent Advances in Intrusion Detection. Springer Berlin Heidelberg, 2000.
188. Ghosh, Anup K., and Aaron Schwartzbard. *"A Study in Using Neural Networks for Anomaly and Misuse Detection."* USENIX Security. 1999.
189. Ghosh, Anup K., Aaron Schwartzbard, and Michael Schatz. *"Learning Program Behavior Profiles for Intrusion Detection."* Workshop on Intrusion Detection and Network Monitoring. Vol. 51462. 1999.
190. Elman, Jeffrey L. *"Finding structure in time."* Cognitive science 14.2 (1990): 179-211.
191. Hosmer, Hilary H. *"Security is fuzzy!: applying the fuzzy logic paradigm to the multipolicy paradigm."* Proceedings on the 1992-1993 workshop on New security paradigms. ACM, 1993.
192. Bridges, Susan M., and Rayford B. Vaughn. *"Fuzzy data mining and genetic algorithms applied to intrusion detection."* Proceedings of 12th Annual Canadian Information Technology Security Symposium. 2000.
193. Dickerson, John E., and Julie A. Dickerson. *"Fuzzy network profiling for intrusion detection."* Fuzzy Information Processing Society, 2000. NAFIPS. 19th International Conference of the North American. IEEE, 2000.
194. Li, Wei. *"Using genetic algorithm for network intrusion detection."* Proceedings of the United States Department of Energy Cyber Security Group 1 (2004): 1-8.
195. Pillai, M. M., Jan HP Eloff, and H. S. Venter. *"An approach to implement a network intrusion detection system using genetic algorithms."* Proceedings of the 2004 annual research conference of the South African institute of computer scientists and information technologists on IT research in developing countries. South African Institute for Computer Scientists and Information Technologists, 2004.
196. Gomez, Jonatan, and Dipankar Dasgupta. *"Evolving fuzzy classifiers for intrusion detection."* Proceedings of the 2002 IEEE Workshop on Information Assurance. Vol. 6. No. 3. New York: IEEE Computer Press, 2002.
197. Jain, Anil K., and Richard C. Dubes. *"Algorithms for clustering data."* Prentice-Hall, Inc., 1988.
198. Tan, Pang-Ning. *"Introduction to data mining."* Pearson Education India, 2006.
199. Ramaswamy, Sridhar, Rajeev Rastogi, and Kyuseok Shim. *"Efficient algorithms for mining outliers from large data sets."* ACM Sigmod Record. Vol. 29. No. 2. ACM, 2000.
200. Sequeira, Karlton, and Mohammed Zaki. *"ADMIT: anomaly-based data mining for intrusions."* Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2002.
201. Ester, Martin, et al. *"A density-based algorithm for discovering clusters in large spatial databases with noise."* Kdd. Vol. 96. No. 34. 1996.
202. Guha, Sudipto, Rajeev Rastogi, and Kyuseok Shim. *"ROCK: A robust clustering algorithm for categorical attributes."* Information systems 25.5 (2000): 345-366.
203. Chandola, Varun, et al. *"Data mining for cyber security."* Data Warehousing and Data Mining Techniques for Computer Security 20 (2006).
204. Smith, Rasheda, et al. *"Clustering approaches for anomaly based intrusion detection."* Proceedings of intelligent engineering systems through artificial neural networks (2002): 579-584.

205. Jiang, Mon-Fong, Shian-Shyong Tseng, and Chih-Ming Su. *"Two-phase clustering process for outliers detection."* Pattern recognition letters 22.6 (2001): 691-700.
206. Mahoney, Matthew V., and Philip K. Chan. *"Learning rules for anomaly detection of hostile network traffic."* Data Mining, 2003. ICDM 2003. Third IEEE International Conference on. IEEE, 2003.
207. Eskin, Eleazar, et al. *"A geometric framework for unsupervised anomaly detection."* Applications of data mining in computer security. Springer US, 2002. 77-101.
208. Pires, A., and Carla Santos-Pereira. *"Using clustering and robust estimators to detect outliers in multivariate data."* Proceedings of the International Conference on Robust Statistics. 2005.
209. Otey, M., et al. *"Towards nic-based intrusion detection."* Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2003.
210. He, Zengyou, Xiaofei Xu, and Shengchun Deng. *"Discovering cluster-based local outliers."* Pattern Recognition Letters 24.9 (2003): 1641-1650.
211. Gogoi, Prasanta, et al. *"A survey of outlier detection methods in network anomaly identification."* The Computer Journal (2011): bxr026.
212. Otey, Matthew Eric, Amol Ghoting, and Srinivasan Parthasarathy. *"Fast distributed outlier detection in mixed-attribute data sets."* Data mining and knowledge discovery 12.2-3 (2006): 203-228.
213. Bhuyan, Monowar H., D. K. Bhattacharyya, and Jugal K. Kalita. *"NADO: network anomaly detection using outlier approach."* Proceedings of the 2011 International Conference on Communication, Computing & Security. ACM, 2011.
214. Verleysen, Michel. *"Learning high-dimensional data."* Nato Science Series Sub Series III Computer And Systems Sciences 186 (2003): 141-162.
215. James, Gareth, Daniela Witten, and Trevor Hastie. *"An Introduction to Statistical Learning: With Applications in R."* (2014).
216. Aggarwal, Charu C., and Philip S. Yu. *"Outlier detection for high dimensional data."* ACM Sigmod Record. Vol. 30. No. 2. ACM, 2001.
217. Blum, Avrim L., and Pat Langley. *"Selection of relevant features and examples in machine learning."* Artificial intelligence 97.1 (1997): 245-271.
218. Pena, Jose Manuel, et al. *"Dimensionality reduction in unsupervised learning of conditional Gaussian networks."* IEEE Transactions on Pattern Analysis and Machine Intelligence 23.6 (2001): 590-603.
219. Yu, Lei, and Huan Liu. *"Feature selection for high-dimensional data: A fast correlation-based filter solution."* ICML. Vol. 3. 2003.
220. Chandrashekar, Girish, and Ferat Sahin. *"A survey on feature selection methods."* Computers & Electrical Engineering 40.1 (2014): 16-28.
221. Shardlow, Matthew. *"An analysis of feature selection techniques."* The University of Manchester (2016).
222. Kohavi, Ron, and George H. John. *"Wrappers for feature subset selection."* Artificial intelligence 97.1-2 (1997): 273-324.
223. Kumar, Vipin. *"Chapman & Hall/CRC Data Mining and Knowledge Discovery Series."* (2010).
224. Dua, Sumeet, and Xian Du. *"Data mining and machine learning in cybersecurity."* CRC press, 2016.
225. Berkhin, Pavel. *"A survey of clustering data mining techniques."* Grouping multidimensional data. Springer Berlin Heidelberg, 2006. 25-71.

226. Fahad, Adil, et al. *"A survey of clustering algorithms for big data: Taxonomy and empirical analysis."* IEEE transactions on emerging topics in computing 2.3 (2014): 267-279.
227. Park, Hae-Sang, and Chi-Hyuck Jun. *"A simple and fast algorithm for K-medoids clustering."* Expert systems with applications 36.2 (2009): 3336-3341.
228. Huang, Zhexue. *"Extensions to the k-means algorithm for clustering large data sets with categorical values."* Data mining and knowledge discovery 2.3 (1998): 283-304.
229. Kaufman, Leonard, and Peter J. Rousseeuw. *"Finding groups in data: an introduction to cluster analysis."* Vol. 344. John Wiley & Sons, 2009.
230. Ng, Raymond T., and Jiawei Han. *"CLARANS: A method for clustering objects for spatial data mining."* IEEE transactions on knowledge and data engineering 14.5 (2002): 1003-1016.
231. Xie, Xuanli Lisa, and Gerardo Beni. *"A validity measure for fuzzy clustering."* IEEE Transactions on pattern analysis and machine intelligence 13.8 (1991): 841-847.
232. Zhang, Tian, Raghu Ramakrishnan, and Miron Livny. *"BIRCH: an efficient data clustering method for very large databases."* ACM Sigmod Record. Vol. 25. No. 2. ACM, 1996.
233. Guha, Sudipto, Rajeev Rastogi, and Kyuseok Shim. *"CURE: an efficient clustering algorithm for large databases."* ACM Sigmod Record. Vol. 27. No. 2. ACM, 1998.
234. Guha, Sudipto, Rajeev Rastogi, and Kyuseok Shim. *"ROCK: A robust clustering algorithm for categorical attributes."* Information systems 25.5 (2000): 345-366.
235. Karypis, George, Eui-Hong Han, and Vipin Kumar. *"Chameleon: Hierarchical clustering using dynamic modeling."* Computer 32.8 (1999): 68-75.
236. Ankerst, Mihael, et al. *"OPTICS: ordering points to identify the clustering structure."* ACM Sigmod record. Vol. 28. No. 2. ACM, 1999.
237. Xu, X., Ester, M., Kriegel, H. P., & Sander, J. *"A nonparametric clustering algorithm for knowledge discovery in large spatial databases."* Proc. of the Intl. Conf. on Data Engineering (ICDE'98). 1998.
238. Hinneburg, Alexander, and Daniel A. Keim. *"An efficient approach to clustering in large multimedia databases with noise."* KDD. Vol. 98. 1998.
239. Sheikholeslami, Gholamhosein, Surojit Chatterjee, and Aidong Zhang. *"Wavecluster: A multi-resolution clustering approach for very large spatial databases."* VLDB. Vol. 98. 1998.
240. Wang, Wei, Jiong Yang, and Richard Muntz. *"STING: A statistical information grid approach to spatial data mining."* VLDB. Vol. 97. 1997.
241. Fraley, Chris, and Adrian E. Raftery. *"MCLUST: Software for model-based cluster analysis."* Journal of classification 16.2 (1999): 297-306.
242. Fraley, Chris, and Adrian E. Raftery. *"MCLUST version 3: an R package for normal mixture modeling and model-based clustering."* WASHINGTON UNIV SEATTLE DEPT OF STATISTICS, 2006.
243. Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. *"Maximum likelihood from incomplete data via the EM algorithm."* Journal of the royal statistical society. Series B (methodological) (1977): 1-38.
244. Fisher, Douglas H. *"Knowledge acquisition via incremental conceptual clustering."* Machine learning 2.2 (1987): 139-172.
245. Gennari, John H., Pat Langley, and Doug Fisher. "*Models of incremental concept formation."* Artificial intelligence 40.1-3 (1989): 11-61.
246. Donoho, David L. *"High-dimensional data analysis: The curses and blessings of dimensionality."* AMS Math Challenges Lecture 1 (2000): 32.

247. Assent, Ira. *"Clustering high dimensional data."* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2.4 (2012): 340-350.
248. Kwak, Nojun, and Chong-Ho Choi. *"Input feature selection for classification problems."* IEEE Transactions on Neural Networks 13.1 (2002): 143-159.
249. Swiniarski, Roman W., and Andrzej Skowron. *"Rough set methods in feature selection and recognition."* Pattern recognition letters 24.6 (2003): 833-849.
250. Kriegel, Hans-Peter, Peer Kröger, and Arthur Zimek. *"Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering."* ACM Transactions on Knowledge Discovery from Data (TKDD) 3.1 (2009): 1.
251. Kriegel, Hans-Peter, Peer Kröger, and Arthur Zimek. *"Subspace clustering."* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2.4 (2012): 351-364.
252. Aggarwal, Charu C., et al. *"Fast algorithms for projected clustering."* ACM SIGMoD Record. Vol. 28. No. 2. ACM, 1999.
253. Bansal, Nikhil, Avrim Blum, and Shuchi Chawla. *"Correlation clustering."* Foundations of Computer Science, 2002. Proceedings. The 43rd Annual IEEE Symposium on. IEEE, 2002.
254. Madeira, Sara C., and Arlindo L. Oliveira. *"Biclustering algorithms for biological data analysis: a survey."* IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 1.1 (2004): 24-45.
255. Liu, Ling, and M. Tamer Özsu. "Encyclopedia of database systems." Vol. 6. Berlin, Heidelberg, Germany: Springer (2009): 2873-2875.
256. Hartigan, John A. *"Direct clustering of a data matrix."* Journal of the american statistical association 67.337 (1972): 123-129.
257. Cheng, Yizong, and George M. Church. *"Biclustering of expression data."* Ismb. Vol. 8. No. 2000. 2000.
258. Wang, Haixun, et al. *"Clustering by pattern similarity in large data sets."* Proceedings of the 2002 ACM SIGMOD international conference on Management of data. ACM, 2002.
259. Pei, Jian, et al. *"Maple: A fast algorithm for maximal pattern-based clustering."* Data Mining, 2003. ICDM 2003. Third IEEE International Conference on. IEEE, 2003.
260. Cho, Hyuk, et al. *"Minimum sum-squared residue co-clustering of gene expression data."* Proceedings of the 2004 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2004.
261. Liu, Jinze, and Wei Wang. *"Op-cluster: Clustering by tendency in high dimensional space."* Data Mining, 2003. ICDM 2003. Third IEEE International Conference on. IEEE, 2003.
262. Kriegel, Hans-Peter, and Arthur Zimek. *"Subspace clustering, ensemble clustering, alternative clustering, multiview clustering: what can we learn from each other."* Proceedings of MultiClustKDD (2010).
263. Zhang, Qiang, et al. *"A top-down search grid based algorithm for fast subspace clustering."* Machine Learning and Cybernetics, 2008 International Conference on. Vol. 1. IEEE, 2008.
264. Aggarwal, Charu C., and Philip S. Yu. *"Finding generalized projected clusters in high dimensional spaces."* Vol. 29. No. 2. ACM, 2000.
265. Woo, Kyoung-Gu, et al. *"FINDIT: a fast and intelligent subspace clustering algorithm using dimension voting."* Information and Software Technology 46.4 (2004): 255-271.
266. Yang, Jiong, et al. "*δ-clusters: capturing subspace correlation in a large data set."* Data Engineering, 2002. Proceedings. 18th International Conference on. IEEE, 2002.

267. Friedman, Jerome H., and Jacqueline J. Meulman. *"Clustering objects on subsets of attributes (with discussion)."* Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66.4 (2004): 815-849.
268. Vahdat, Ali, Malcolm Heywood, and Nur Zincir-Heywood. *"Bottom-up evolutionary subspace clustering."* Evolutionary Computation (CEC), 2010 IEEE Congress on. IEEE, 2010.
269. Parsons, Lance, Ehtesham Haque, and Huan Liu. *"Subspace clustering for high dimensional data: a review."* Acm Sigkdd Explorations Newsletter 6.1 (2004): 90-105.
270. Cheng, Chun-Hung, Ada Waichee Fu, and Yi Zhang. *"Entropy-based subspace clustering for mining numerical data."* Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 1999.
271. Goil, Sanjay, Harsha Nagesh, and Alok Choudhary. *"MAFIA: Efficient and scalable subspace clustering for very large data sets."* Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 1999.
272. Chang, Jae-Woo, and Du-Seok Jin. *"A new cell-based clustering method for large, high-dimensional data in data mining applications."* Proceedings of the 2002 ACM symposium on Applied computing. ACM, 2002.
273. Procopiuc, Cecilia M., et al. *"A Monte Carlo algorithm for fast projective clustering."* Proceedings of the 2002 ACM SIGMOD international conference on Management of data. ACM, 2002.
274. Liu, Bing, Yiyuan Xia, and Philip S. Yu. *"Clustering through decision tree construction."* Proceedings of the ninth international conference on Information and knowledge management. ACM, 2000.
275. Yip, Kevin Y., M. K. Ng, and D. W. Cheung. *"A review on projected clustering algorithms."* International Journal of Applied Mathematics 13.1 (2003): 35-48.
276. Sembiring, Rahmat Widia, Jasni Mohamad Zain, and Abdullah Embong. *"Clustering high dimensional data using subspace and projected clustering algorithms."* arXiv preprint arXiv:1009.0384 (2010).
277. Yip, K. P., David W. Cheung, and Michael K. Ng. *"On discovery of extremely low-dimensional clusters using semi-supervised projected clustering."* Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on. IEEE, 2005.
278. Bohm, Christian, et al. *"Density connected clustering with local subspace preferences."* Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on. IEEE, 2004.
279. Yiu, Man Lung, and Nikos Mamoulis. *"Frequent-pattern based iterative projected clustering."* Data Mining, 2003. ICDM 2003. Third IEEE International Conference on. IEEE, 2003.
280. Yiu, Man Lung, and Nikos Mamoulis. *"Iterative projected clustering by subspace mining."* IEEE Transactions on Knowledge and Data Engineering 17.2 (2005): 176-189.
281. Moise, Gabriela, Jorg Sander, and Martin Ester. *"P3C: A robust projected clustering algorithm."* Data Mining, 2006. ICDM'06. Sixth International Conference on. IEEE, 2006.
282. Moise, Gabriela, Jörg Sander, and Martin Ester. *"Robust projected clustering."* Knowledge and Information Systems 14.3 (2008): 273-298.
283. Jing, Liping, Michael K. Ng, and Joshua Zhexue Huang. *"An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data."* IEEE Transactions on knowledge and data engineering 19.8 (2007).

284. Domeniconi, Carlotta, et al. *"Subspace clustering of high dimensional data."* Proceedings of the 2004 SIAM international conference on data mining. Society for Industrial and Applied Mathematics, 2004.
285. Huang, Joshua Zhexue, et al. *"Automated variable weighting in k-means type clustering."* IEEE Transactions on Pattern Analysis and Machine Intelligence 27.5 (2005): 657-668.
286. Bouveyron, Charles, Stéphane Girard, and Cordelia Schmid. "High-dimensional data clustering." Computational Statistics & Data Analysis 52.1 (2007): 502-519.
287. Domeniconi, Carlotta, et al. *"Locally adaptive metrics for clustering high dimensional data."* Data Mining and Knowledge Discovery 14.1 (2007): 63-97.
288. Lu, Yanping, et al. *"Particle swarm optimizer for variable weighting in clustering high-dimensional data."* Swarm Intelligence Symposium, 2009. SIS'09. IEEE. IEEE, 2009.
289. Assent, Ira, et al. *"DUSC: Dimensionality unbiased subspace clustering."* Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on. IEEE, 2007.
290. Liu, Guimei, et al. *"Distance based subspace clustering with flexible dimension partitioning."* Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on. IEEE, 2007.
291. Achtert, Elke, et al. *"Detection and visualization of subspace cluster hierarchies."* Advances in databases: Concepts, systems and applications (2007): 152-163.
292. Yip, Kevin Y., David W. Cheung, and Michael K. Ng. *"Harp: A practical projected clustering algorithm."* IEEE Transactions on knowledge and data engineering 16.11 (2004): 1387-1397.
293. Sequeira, Karlton, and Mohammed Zaki. *"SCHISM: a new approach to interesting subspace mining."* International Journal of Business Intelligence and Data Mining 1.2 (2005): 137-160.
294. Kriegel, H-P., et al. *"A generic framework for efficient subspace clustering of high-dimensional data."* Data Mining, Fifth IEEE International Conference on. IEEE, 2005.
295. Kriegel, Hans-Peter, Peer Kröger, and Arthur Zimek. *"Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering."* ACM Transactions on Knowledge Discovery from Data (TKDD) 3.1 (2009): 1.
296. Aggarwal, Charu C., and Philip S. Yu. *"Finding generalized projected clusters in high dimensional spaces."* Vol. 29. No. 2. ACM, 2000.
297. Böhm, Christian, et al. *"Computing clusters of correlation connected objects."* Proceedings of the 2004 ACM SIGMOD international conference on Management of data. ACM, 2004.
298. Achtert, Elke, et al. *"Robust, complete, and efficient correlation clustering."* Proceedings of the 2007 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2007.
299. Achtert, Elke, et al. *"On exploring complex relationships of correlation clusters."* Scientific and Statistical Database Management, 2007. SSBDM'07. 19th International Conference on. IEEE, 2007.
300. Achtert, Elke, et al. *"Robust clustering in arbitrarily oriented subspaces."* Proceedings of the 2008 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2008.
301. Achtert, Elke, et al. *"Global correlation clustering based on the Hough transform."* Statistical Analysis and Data Mining 1.3 (2008): 111-127.
302. Cordeiro, Robson Leonardo Ferreira, et al. *"Finding Clusters in subspaces of very large, multi-dimensional datasets."* ICDE. 2010.

303. Fischler, Martin A., and Robert C. Bolles. *"Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography."* Communications of the ACM 24.6 (1981): 381-395.
304. Leung, Kingsly, and Christopher Leckie. *"Unsupervised anomaly detection in network intrusion detection using clusters."* Proceedings of the Twenty-eighth Australasian conference on Computer Science-Volume 38. Australian Computer Society, Inc., 2005.
305. Casas, Pedro, Johan Mazel, and Philippe Owezarski. *"Unada: Unsupervised network anomaly detection using sub-space outliers ranking."* NETWORKING 2011 (2011): 40-51.
306. Casas, Pedro, Johan Mazel, and Philippe Owezarski. *"Steps towards autonomous network security: unsupervised detection of network attacks."* New Technologies, Mobility and Security (NTMS), 2011 4th IFIP International Conference on. IEEE, 2011.
307. Ishida, Moriteru, Hiroki Takakura, and Yasuo Okabe. *"High-performance intrusion detection using optigrid clustering and grid-based labelling."* Applications and the Internet (SAINT), 2011 IEEE/IPSJ 11th International Symposium on. IEEE, 2011.